Samuel Rönnqvist

Knowledge-Lean Text Mining

Turku Centre *for* Computer Science

TUCS Dissertations
No 227, December 2017

Knowledge-Lean Text Mining

Samuel Rönnqvist

To be presented, with the permission of the Faculty of Science and Engineering of Åbo Akademi University, for public criticism in Agora, Auditorium XX on December 8, 2017.

Doctoral thesis in Computer Science
Åbo Akademi University
Turku, Finland

2017

Advisors

Docent Tomas Eklund
Faculty of Science and Engineering
Åbo Akademi University
Finland

Professor emerita Barbro Back
Faculty of Science and Engineering
Åbo Akademi University
Finland

Reviewers

Professor Chris Biemann
Department of Informatics – Language Technology
University of Hamburg
Germany

Professor Uzay Kaymak
Industrial Engineering & Innovation Sciences – Information Systems
Eindhoven University of Technology
The Netherlands

Opponent

Professor Krista Lagus
Faculty of Social Sciences – Digital Social Science
University of Helsinki
Finland

ISBN 978-952-12-3622-8
ISSN 1239-1883

Abstract

This thesis explores the process of introducing text mining to new areas of application, which involves both defining appropriate types of analysis and often designing appropriate computational methods to support the analysis. Targeted toward a particular use, text mining resources tend to become highly specialized and require considerable efforts in development. The thesis addresses the question of what computational methods can serve practical text analysis needs, while avoiding costly and narrow development of linguistic resources.

Relying on machine learning and visualization, this knowledge-lean approach assumes minimal encoding of prior knowledge into resources, which is essential in entering uncharted text mining territory, that is, areas too new or too marginal to be well served by traditional text mining approaches. Knowledge-lean text mining is explored within the domain of systemic financial risk, where few text mining efforts have previously been pursued.

Without the support of existing linguistic resources for the task, unsupervised and data-driven methods play a key role in providing flexible means for text analysis. The central theme of representation learning is studied also in the context of fully knowledge-free, domain-independent topic modeling and linguistically resource-lean discourse structure parsing for the refinement of text mining results.

The research has been able to establish the value of knowledge-lean text mining, by exploring the use of text as a source of information for systemic risk analytics. Furthermore, the work on discourse parsing has shown that competitive – and in some cases state-of-the-art – performance can be achieved without relying on explicit encoding of linguistic knowledge.

Sammanfattning

Kunskapssnål textanalytik

Denna avhandling behandlar algoritmisk textanalys, *textanalytik* (även *textutvinning*; eng. *text mining, text analytics*), och hur man kan gå till väga för att introducera den till nya tillämpningsområden. Att ta sig an nya tillämpningar innebär att lämpliga typer av analys måste definieras, samt ofta att lämpliga algoritmiska metoder måste utvecklas för att understödja analysen. När den algortimiska textanalysen inriktas på en specifik tillämpning tenderar resurserna som utnyttjas att bli mycket specialiserade och kräva avsevärt arbete att utveckla. Avhandlingen belyser frågan kring hur algoritmiska metoder kan utformas för att understödja praktiska textanalysbehov, medan de gör det möjligt att kringgå behovet av kostsam och opraktiskt smal utveckling av lingvistiska resurser.

Ett sådant kunskapssnålt tillvägagångssätt, som förlitar sig på maskininlärning och visualisering, förutsätter endast till en minimal grad kodande av kunskap i form av resurser. Kunskapssnålheten är väsentlig för att lättare kunna bryta ny mark inom datautvinning på text, det vill säga utforska tillämpningsområden som är för nya eller marginella för att understödjas väl av traditionella tillvägagångssätt. Kunskapssnål textanalytik undersöks främst inom domänen finansiell systemrisk, där få tidigare försök till automatiserad textanalys har gjorts.

Utan existerande lingvistiska resurser som stöd för uppgiften spelar oövervakade och datadrivna metoder en nyckelroll som flexibla medel för textanalys. Representationsinlärning är ett centralt tema som undersöks i kontexten av helt kunskapsfri och domänoberoende modellering av ämnen, samt lingvistiskt resurssnål analys av diskursstruktur för att förfina informationen som utvinns ur text.

Forskningen har kunnat påvisa värdet av kunskapssnål textanalytik genom att utforska användningen av text som en ny informationskälla inom analytik för studie av systemrisk. Dessutom har arbetet kring analys av diskursstruktur visat att toppresultat kan uppnås utan att man förlitar sig på explicit kodande av lingvistisk kunskap.

Acknowledgements

As this enduring project is coming to an end, I am grateful for all the support I have received. It has been a largely enjoyable experience, being able to channel my curiosity and creativity into research, and grow with the task. I would like to thank the people who have helped provide the context for me to work in, offered advice, inspiration and their time.

I thank my advisors Tomas Eklund and Barbro Back for taking on the responsibility and accompanying me in my pursuit of a doctoral degree. I extend my gratitude toward my fellow Ph.D. students and co-authors of the Data Mining and Knowledge Management Laboratory and the department: Annika Holmbom, Henri Korvela, Hongyan Liu, Peter Sarlin, Zhiyuan Yao and Xiaolu Wang. You have all made the journey less lonely and more interesting. My collaboration with Peter has been particularly productive and educational. I also thank the rest of the IT department for having made it such a welcoming environment.

I would like to acknowledge the financial support that has allowed me the freedom to explore, learn about and become experienced in the topics I have found most interesting along the road. This has been provided by the Graduate School of Åbo Akademi University, Turku Centre for Computer Science Graduate Programme, Goethe University Frankfurt, the Nokia Foundation, the Hans Bang Foundation and the Åbo Akademi University Foundation.

I am thankful to Christian Chiarcos for kindly inviting me to spend a semester at the Applied Computational Linguistics Laboratory at Goethe University Frankfurt. The direct insight into the field of computational linguistics has been an enriching experience. A warm thank you to Niko Schenk, for the many engaging and fruitful discussions and hard work that have taken us from MLSS via FFM to the ACL.

Filip Ginter at the University of Turku has been a true inspiration, working with him convinced me to fully settle on the path of text mining. Thank you Filip and the rest of the Turku NLP Group whose machine learning and NLP mastery has provided me much valued guidance.

Another important part of this journey has been the numerous and lively discussions that often went far beyond anyone's particular research focus, but certainly not our interests. Among other things, I fondly remember

the SEMPRE meetings, coffee with the embedded systems guys, information systems afterworks, and many conferences. Thank you Espen Suenson, Henrik Nyman, Piia Hirkman, Marta Olszewska, Natalia Díaz Rodríguez, Magnus Westerlund, Johan Ersfolk, Kalle Rönnholm, Mats Neovius, Bassam Mokbel, Rob Golan and Luana Micallef. I have also enjoyed working outside a strictly academic setting and would particularly like to extend my thanks to John Kronberg and Mikael Sand.

I might not have ended up a computer scientist, had it not been for Wictor Lund and Simon Holmbacka, with whom I started programming in the late 90s. Tinkering with hardware and software throughout our teenage years, in a wonderfully spurring, creative dynamic, we no doubt paved the way that led us here, as both colleagues and friends. Tack!

I want to extend my sincere gratitude to my parents, Margareta and Ralph Rönnqvist, as well as my siblings Hanna and Simon, whom I have a lot to thank for, not the least for fostering the curiosity and persistence so vital for completing a doctorate (and for bringing home a computer and uplink in the early 90s). Lastly, I extend my most heartfelt gratitude to my wife Miriam for continuing to broaden my horizons, and for having been so considerate and patient throughout this process. Our daughter Selma brings sunshine every day, and she offers truly new perspectives on life, learning and language. Vielen Dank!

List of original publications

I Samuel Rönnqvist and Peter Sarlin. "Bank networks from text: Interrelations, centrality and determinants." *Quantitative Finance*, 15(10), 1619–1635. 2015. doi:10.1080/14697688.2015.1071076

II Samuel Rönnqvist and Peter Sarlin. "Bank distress in the news: Describing events through deep learning." *Neurocomputing*, 264, 57–70. 2017. doi:10.1016/j.neucom.2016.12.110

III Samuel Rönnqvist, Xiaolu Wang and Peter Sarlin. "Interactive visual exploration of topic models using graphs." In *Proceedings of the Eurographics Conference on Visualization (EuroVis)*, 2014.

IV Samuel Rönnqvist. "Exploratory topic modeling with distributional semantics." In: Fromont E., De Bie T., van Leeuwen M. (eds), *Advances in Intelligent Data Analysis XIV*. Lecture Notes in Computer Science, 9385, 241–252. 2015. doi:10.1007/978-3-319-24465-5_21

V Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Rönnqvist, Evgeny A. Stepanov and Giuseppe Riccardi. "Do we really need all those rich linguistic features? A neural network-based approach to implicit sense labeling." In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL) – Shared Task*, 2016. doi:10.18653/v1/k16-2005

VI Samuel Rönnqvist, Niko Schenk and Christian Chiarcos. "A recurrent neural model with attention for the recognition of Chinese implicit discourse relations." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. doi:10.18653/v1/P17-2040

List of other co-authored publications

1. Samuel Rönnqvist and Peter Sarlin. "Detect & Describe: Deep learning of bank stress in the news." In *Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2015.

2. Samuel Rönnqvist and Peter Sarlin. "Identifying bank stress by deep learning of news." *Machine Learning Reports*, 3, 112–113. 2015.

3. Samuel Rönnqvist and Peter Sarlin. "Alluvial SOTM: Visualizing transitions and changes in cluster structure of the Self-Organizing Time Map." In *Proceedings of the Eurographics Conference on Visualization (EuroVis)*, 2014.

4. Samuel Rönnqvist and Peter Sarlin. "From text to bank interrelation maps". In *Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, 2014.

5. Annika H. Holmbom, Samuel Rönnqvist, Peter Sarlin, Tomas Eklund and Barbro Back. "Green vs. non-green customer behavior: A Self-Organizing Time Map over greenness." In *Proceedings of the 13th IEEE International Conference on Data Mining Workshops (ICDMW)*, 2013.

6. Peter Sarlin and Samuel Rönnqvist. "Cluster coloring of the Self-Organizing Map: An information visualization perspective." In *Proceedings of the 17th International Conference on Information Visualisation*, 2013.

7. Sofie Van Landeghem, Kai Hakala, Samuel Rönnqvist, Tapio Salakoski, Yves Van de Peer and Filip Ginter. "Exploring biomolecular literature with EVEX: Connecting genes through events, homology, and indirect associations." *Advances in Bioinformatics*, vol. 2012, Article ID 582765, 12 pages. 2012.

Contents

Chapter 1

Introduction

"Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?"

– T.S. Eliot (1888-1965) [66]

1.1 Context

Innumerable by now are the opening statements of research papers alluding to the overwhelming and evergrowing amount of information that surrounds us, followed by the logical conclusion to study computational methods that can help us analyze, organize, filter and make sense of it. Our environment is indeed information rich, and making sense of one's environment is a most basic human urge, in fact, a defining property of intelligence [87].

1.1.1 Machine thinking

As our mental capacities are restricted by biology, it is tempting to look for ways to extend our information processing capabilities beyond our bodies. The first steps in science to near the age old dream of creating artificial intelligent things or beings were taken as Aristotle embarked upon the quest for mechanical thinking, work on logic which was later continued by Llull, Leibniz, Boole and others in the last millennium [157]. By 1943, McCulloch and Pitts [137], inspired by neural information processing in the brain, laid the foundations for artificial neural networks, which later have established an important paradigm for machine intelligence. In 1950, interested in the question of whether machines can think, Turing proposed the *imitation game* as a test for human-like intelligent behavior in machines [203]. This decade saw a sprawling interest in *artificial intelligence* (AI), made possible by the

advent of general-purpose digital computers, and the term itself was coined by John McCarthy in 1956 [157]. That year, the cybernetic concept of *intelligence amplification* [8] by computers was introduced as well, followed in the 1960s by similar work on the integration of computational information processing to support and augment the human intellect [126, 67].

Much of the artificial intelligence research in the later half of the century dealt with symbolic knowledge representations and symbol processing, whereas neural information processing and statistical decision-making approaches developed as parallel paradigms [157]. The symbol-based, classical artificial intelligence was met with early enthusiasm but proved unable to move beyond very narrow applications toward the dream of general artificial intelligence. According to critics, such as Dreyfus [64], systems based on the encoding of rules and facts would not be able to scale, nor capture commonsense or tacit knowledge necessary for more general reasoning about the world, while learning-based systems of the time were too simplistic in their own right. Neural network concepts stemming from the 20^{th} century have caused a renaissance in artificial intelligence in the last few years, which largely has been driven by the increase in available computing power and data [184].

Hawkins and Blakeslee [87] suggest that defining intelligence by behavior is an impediment to understanding what intelligence is, and that intelligence is defined more accurately in terms of prediction. Accordingly, classical artificial intelligence, which has focused extensively on producing human-like behavior in narrow tasks, rather than seeking to understand the nature of intelligence, has had problems in scaling. *Machine learning* (Section 2.2) fits better with their view, namely that intelligence in essence is the ability to learn from experience in order to predict on sequential input, e.g., predicting the next word in a sentence you hear.[1] Predictive sequence modeling is a corner stone in many recent advances in neural-network-based artificial intelligence, neural language models not the least (see Sections 2.2.2 and 2.2.3).

While the field of artificial intelligence has concentrated very much on achieving autonomous intelligent behavior, the notion set forth by intelligence amplification, that machine intelligence often be best posed to serve human intelligence rather than acting independently, deserves more recognition and attention. Many of the most interesting and useful applications of machine intelligence arguably are those that integrate well with people in order to support their cognitive tasks, such as data analysis. This thesis rests upon the artificial intelligence tradition in its focus on processing human

[1] As they describe, the brain constantly makes predictions based on the sequence of inputs it is receiving, and based on what it has learned from previous experiences, and that is the essence of intelligence and understanding the environment. Intelligent behavior, i.e., acting intelligently in an environment, follows from understanding.

language and the use of machine learning (e.g., artificial neural networks; Section 2.2.2), as well as upon intelligence amplification in the incarnation of *visual analytics* (see Section 2.3.3).

1.1.2 From data mining to text mining

Data mining, a concept that started to appear by the end of the 1980s [169], is as a field of study focused on making sense of an environment of abundant data. Also referred to as *knowledge discovery in databases* (KDD), this field often takes a rather pragmatic aim, e.g., on practical business problems. While the classical application of statistics often followed a process of theory-backed hypothesis formulation, followed by gathering of data and hypothesis testing, the new abundance of stored data turned this process on its head [90, 61]. Statistical methods remained important tools in data mining along with a growing body of algorithmic and machine learning methods. Nevertheless, databases were used increasingly to discover unanticipated patterns and trends without preformulated hypotheses.

The data mining community had also laid its eyes on the growing amount of stored text, and saw text data mining, or *text mining*, as a natural extension of its field. Early efforts in the 1990s addressed the challenge of mining text [169], which they viewed as unstructured data. Processing human language text is, however, as difficult as it is enticing. Human language is ridden with ambiguity and high variability[2] that makes it difficult to decode from surface form to structured representations. Nonetheless, it encodes rich and valuable expressive detail in the form of facts, ideas and opinions, i.e., expressions of our understanding of the world. Language is intimately bound to intelligence, as Turing acknowledges by making language understanding and generation a hallmark in his test for machine intelligence [203]. Language plays a central role in making sense of the world, as is elaborated on in Section 2.1.

As Hearst [88] points out, by the turn of the century, text mining had not yet been very successful in delivering on its hype. The anticipation had partly been fueled by the exploding availability of text data on the web, but text mining turned out not to be as straightforward an extension of data mining as expected. In her analysis, Hearst continues to suggest that data mining and text mining actually are more different than many recognize, that it is not simply data mining on text data as usual. Despite its metaphorical name, data mining would be about finding valuable patterns in aggregates of data, rather than extracting valuable nuggets of data. She

[2]Language variability translates into *sparsity* in data. For instance, word frequencies follow a scale-free distribution (see Section 3.2.2), meaning that text corpora have long tails of very infrequent words in their distributions, which constitutes a challenge to empirical analysis.

argues that text mining is closer than data mining to the analogy, in its focus on extracting unanticipated but specific pieces of information.

Text is fundamentally different to other types of data in that text is meaningful at the local level (e.g., phrases, sentences), whereas other data usually are not interesting as individual data points per se, but only in aggregate or in relation. As text mining developed, other views on what it entails emerged; Hotho et al. [101] discuss alternative perspectives. On the one hand, in line with Hearst's opinion, text mining can be seen as information extraction at the local level, i.e., extraction of entities, attributes, relations, etc. On the other hand, text mining is indeed treated as data mining of unstructured data, where natural language processing tools serve to preprocess text in order to introduce structure, before the application of common data mining techniques at the aggregate level. In this respect, text mining may refer to an extended process of finding valuable information in text data, beyond the application of information extraction techniques. Since the turn of the century, language technology has advanced considerably and provided a range of sophisticated tools (see Section 2.1.3), thus, text mining is not just trivially applied data mining on text data anymore.

Hearst moreover compares the search for patterns in data, the general aim of data mining, to corpus-based computational linguistics, which likewise concerns modeling of patterns across data sets. However, she differentiates by stating that text mining is meant to "tell us something about the world, outside of the text collection itself", whereas "computational linguistics applications tell us about how to improve language analysis" [88]. Computational linguistics as a field supports development of *natural language processing* (NLP) tools, which help structure text for mining.

1.1.3 Knowledge intensive vs. knowledge lean

An overall trend that can be observed in the development of text mining systems is from general (e.g., [185, 2]) toward more specific designs, which increasingly involve tools and resources that may be specific to *language*, *domain of application* and often *task*. This specialization is motivated by domain needs, as a focus on more narrowly defined information better serves to answer relevant questions. This is the case for instance in biomedical text mining, where a prominent focus has been on the extraction of detailed relations involving genes, diseases, etc. (cf. [3]), and in financial text mining, e.g., in the extraction of events from news for decision support and trading (cf. [98]). A problem with the approach is that it is costly, as it requires substantial manual effort in encoding knowledge through the creation of tools and linguistic (knowledge) resources. In particular, the issue resides in that narrowly designed and poorly generalizing resources provide little possibility for reuse, beyond a single task, domain/sublanguage or language. Reusabil-

ity and generalizability of resources are two important factors influencing how cost of development translates into utility of a text mining system.

Tools for natural language processing, while extensively reliant on machine learning, have traditionally also been developed with much manual effort. Human input either comes in the form of annotation of training data, which encodes linguistic knowledge *implicitly*, or in the form of other manual engineering of features and resources, which encodes knowledge *explicitly*. Due to the symbolic nature of text, this approach is confined to the limited coverage of the crafted resources, and without any data-driven component for inferring representations, it is unable to generalize. This presents a problem, as the sparsity of text data makes it likely to encounter previously unseen instances.

Natural language processing that makes use of encoded knowledge, especially into resources such as dictionaries, ontologies and other knowledge bases, is referred to as *knowledge based* [132]. The use of knowledge bases has its roots in *knowledge-based systems* [1], classical artificial intelligence and rule-based processing. A contrasting paradigm relies on statistical and machine learning methods for data-driven inference. Biemann describes the early rule-based, introspection-driven approaches to linguistic processing as unable to scale, compared to empiricist approaches that grew popular as more machine-readable text and processing power became available [25]. Some systems avoid knowledge bases and annotated corpora altogether, and are often referred to as *knowledge free* (see, e.g., [141, 174]). This approach of knowledge-free and unsupervised modeling of language is described as part of a *structure discovery* paradigm [24, 25]. In the same direction, Banko et al. [11] introduce *open information extraction* by proposing a self-supervised tool for extracting relations between entities, without restricting the types.

In the last few years, a related paradigm set on automatically learning feature representations, *representation learning* (Section 2.2.3), has gained a strong foothold. It offers data-driven discovery of suitable representations of data that are able to very effectively generalize from annotated instances of language use, while relying on unsupervised learning on large amounts of raw text. Under the umbrella term *deep learning*, this approach has achieved strong state-of-the-art performance in many areas of natural language processing, including syntactic parsing and machine translation, as well as other areas of artificial intelligence such as computer vision and speech recognition [189, 184].

The different approaches to language processing may be positioned along a knowledge-intensiveness scale: originating in the strictly knowledge free and extending toward increasing degrees of *knowledge intensiveness*[3] (see

[3]Tomiyama et al. explore the concept in the context of knowledge-based engineering and observe that, while large-scale knowledge bases are useful in engineering applications,

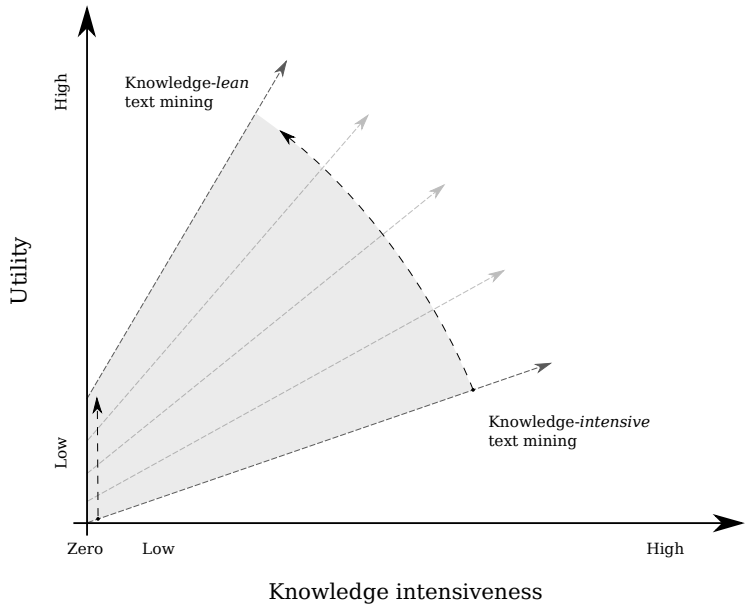


Figure 1.1: Knowledge-lean text mining illustrated in relation to *knowledge intensiveness* and *utility*. The utility of a text mining system is commonly improved through intensified use of knowledge resources (moving rightward), while the efficiency of a given resource (line slope) may be increased by means of data-driven methods. *Knowledge-lean text mining* aims at increasing this efficiency, i.e., achieving practical utility with limited intensiveness, through the appropriate choice of methods.

horizontal axis in Figure 1.1). Knowledge-free natural language processing completely avoids encoded knowledge, whereas a knowledge-intensive system requires large amounts of encoded knowledge. Unsupervised modeling may increase the utility of a given knowledge resource (vertical axis in Figure 1.1) by introducing generalizations and improving its coverage. For instance, representation learning typically is used to generalize from the language patterns in a limited set of annotations, and distant supervision or data augmentation techniques may expand a set of annotations. Such measures may reduce the knowledge intensiveness of a system in practice, as less encoding is needed in order to achieve a certain level of performance and utility.

---

they may 'hard-fail' in the face of unknown situations. [198] They argue that both knowledge intensiveness and flexibility, along with knowledge reuse and sharing, are crucial issues. This seems to support the view I am representing, namely that utility of an application is a function of intensiveness and generalizability of knowledge resources. The closely related concept of *knowledge intensity* has also been studied, e.g., from a knowledge economy perspective [6].

Slightly relaxing the knowledge-free criterion, I define a lightweight approach to language processing that allows for limited use of encoded knowledge as *knowledge lean*.[4] Hence, moving beyond the modeling of language itself to a focus on modeling text content, while still adhering to the philosophy of minimal knowledge intensiveness, this thesis explores the realm of *knowledge-lean text mining* (illustrated in Figure 1.1 by the steeper line, residing on the left hand side). This is studied as a highly data-driven and more flexible alternative to the common knowledge-intensive approaches to text mining.

The sloped lines in Figure 1.1 illustrate the presumed relationship between knowledge intensiveness of linguistic resources and the utility of text mining systems that utilize them.[5] The lower line illustrates the starting point: without data-driven methods that provide generalization, the utility is increased by introducing more encoded knowledge, either in the form of reusable existing resources or by creating new resources, which is costly. However, the aim in this thesis is to introduce computational methods that may provide utility without the need to rely extensively on encoded knowledge.

Sometimes the term *resource lean* and knowledge lean are used interchangeably (cf., e.g., [86, 183]), and should be understood as requiring little linguistic resources. They may both be understood as knowledge resource lean. Resource lean is typically used in a representation learning context, where annotated corpora are used for training, with inferred features only.

1.1.4 Text mining vs. text analytics

During the past decade, the term *analytics* grew in popularity,[6] with a contemporary definition of "systematic computational analysis of data or statistics" (or the information resulting thereof)[7]. Along the same lines, Grimes explains the difference between analysis and analytics:

[4]The term knowledge lean has seen some early although limited use, e.g., by [167] in an NLP context synonymously to knowledge free, and similarly by [103] within AI. In psychology, the concept of *knowledge-lean problems* is more prevalent, being defined as problems that require little or no knowledge to solve (cf., e.g., [207]). The cognitive processing to solve problems are placed on a continuum between knowledge lean and knowledge intensive [154, 125].

[5]The hypothesized relationship has not been empirically tested and is not necessarily linear.

[6]For instance, Google Trends shows a steady rise in interest for analytics as a search term between 2005 and 2011, and it remains an order of magnitude more popular than *data mining* or *big data*. Analytics and the equivalent latin term *analytica* nonetheless have their etymological roots in works of Aristotle on logic and scientific method. According to Google Books Ngram Viewer, the English term analytics appeared in print by 1738, but has seen a rapid exponential increase in frequency only in the last decades.

> "Analysis is an examination of structure, composition, and meaning that provides insight to advance some purpose. Analysis may be heuristic, informal, and/or qualitative. Contrast with analytics, which is algorithmic rather than heuristic. I define analytics as the systematic application of numerical and statistical methods that derive and deliver quantitative information, whether in the form of indicators, tables, or visualizations. Analytics is formal and repeatable." [81]

By this definition, analytics is a special case of the superordinate concept of analysis. The practice of analytics meanwhile encompasses a range of computational tools of varying complexity. For instance, web analytics refers to the gathering of statistics of web usage not necessarily involving further modeling of data, whereas the terms visual analytics, business analytics, advanced analytics and predictive analytics emphasize the use of machine learning in various forms. Analytics appears synonymous to data mining in the sense of a process or even particular techniques used, but it remains rather ambiguous without specification and therefore may carry a more general, less technical connotation. The term analytics tends to prevail in business contexts, whereas data mining still remains the preferred term for instance in the scientific community.[7]

Text analytics, as a specification of the term analytics, is used largely synonymously to text mining, although being rooted in somewhat different traditions and contexts of use [81]. These contexts appear rather separate and few comparisons between text mining and text analytics have been made. Grimes [81] defines text analytics in the same way as Hearst [88] defines text mining, i.e., as algorithmic analysis addressing the information content of text rather than the text itself. Text mining and text analytics tend more toward qualitative analysis than data mining and other types of analytics (particularly predictive analytics), due to the qualitative nature of text, but the distinction is not absolute. For instance, text analytics blending with predictive analytics can be highly quantitative in focus, neither does text mining exclude predictive modeling.

Text mining is the preferred term in this thesis, due to its embedding in a more clearly technical context and continued predominance in science. Nevertheless, the term should be understood as interchangeable in essence to text analytics, both terms alluding to technical aspects as well as the integration with the domain of application. Text analysis is used in a general

[7] According to Google Scholar, the number of publications mentioning text analytics has only started to grow notably since the last 10 years, approaching a rate of 1:8 against text mining by the end of 2016. Google Trends also shows a slow and steady increase in use of the term text analytics, but it remains well below a rate of 1:4 against text mining, which has a similar trajectory.

sense, encompassing text mining, manual analysis of text content as well as linguistic analysis.

1.2 Research focus

This thesis is set against the background of introducing text mining to a new problem area, which typically involves both specifying appropriate analysis objectives and designing computational tools to support the analysis. By the traditional knowledge-intensive approaches to text mining, as the analysis becomes more targeted toward a particular task, the resources tend to become highly specialized and costly to develop. This creates a bottleneck that hinders text mining from being more broadly applied to serve a diverse spectrum of users in developing interesting use cases on their own.

1.2.1 Research question

This thesis addresses the question of *what computational methods can serve practical text analysis needs, and to what extent one can avoid the need for costly and highly specialized linguistic resources*. Costly here implies an emphasis on effort of development, which in combination with a high degree of specialization is problematic. Linguistic resources that are *task*, *domain* or *language-specific*, may not support text mining for new problems. While this may not be a concern in many of the areas where text mining has already been successful, answering this question is paramount in order to stimulate and accelerate exploration of uncharted text mining territory, that is, languages, domains and tasks that may be too marginal or too new to be served by traditional approaches. The question is pursued through the case studies discussed below.

1.2.2 Test cases

A domain where few text mining efforts have been pursued is the study of systemic financial risk and financial stability, despite the potential benefits text mining stands to offer in this area. This makes the domain an ideal case for exploring the above question. Data access and timeliness are particular problems in systemic risk analytics, which is why text may contribute as an important source of information in identifying risks, and, importantly, in understanding them better, thanks to the rich expressive detail of text. The applications relating to this case domain focus particularly on understanding, by means of text mining, the risks that banks may pose to the financial system (presented in Section 4.1).

Resting on a shared methodological basis, this thesis also explores computational means for text analysis with a more general focus, which can offer

exploration of unfamiliar text sources with minimal customization needs, or serve as supporting components for specialized text mining tasks. This work is presented as applications to the tasks of exploring topics in corpora (Section 4.2) and parsing of discourse structure (Section 4.3).

1.2.3 Methodology

The discussed knowledge-free philosophy to natural language processing constitutes a guiding principle for this thesis, at the text mining level, as it focuses on modeling primarily the content matter carried by language. Some forms of text mining can be performed in a fully knowledge-free manner, such as the focus on exploration of topics, but rather quickly it becomes necessary to narrow down the analysis to some degree, and specify some points of reference, in order to achieve more meaningful results. Thus, the work departs from a strict knowledge-free modus operandi and explores knowledge-lean text mining methods. Specifying necessary points of reference may involve limited encoding of linguistic patterns (e.g., for the recognition of named entities), to the extent that the effort is pragmatically justifiable. In some cases, domain-related and linguistic knowledge can be decoupled, which promotes data reuse and makes the method more generally applicable.

The knowledge-lean approach is implemented by the use of data-driven methods, which operate by general rules to extract meaningful structure from text. The algorithms specify general strategies for extracting information, rather than knowledge about the language or domain. The methods concerned can be broadly divided into two types: *relational* and *semantic*. A simple form of relation modeling is first explored that does not rely on machine learning. Then, unsupervised learning is introduced to infer representations of meaning, in the vein of representation learning, as a support for various text mining tasks, as well as for natural language processing.

Finally, as a complement to the scalable, repeatable and quantifiable analysis computational modeling performs, the human, capable of nuanced, versatile and contextualized understanding, is embedded into the analysis process, too. Broadly following the ideas of intelligence amplification, computational and human information processing, or intelligence, are joined through the framework of visual analytics. This involves a particular focus on the role of visual interactive presentation of information as the interface between these two sides. While elaborated in Sections 2.3.3 and 3.1, the aim is, in practical terms, to integrate sophisticated computational modeling and user exploration, in order to balance their respective strengths in data analysis. Considering text understanding, which is very challenging to artificial intelligence, the involvement of the human user is particularly helpful.

Following all the above, this thesis focuses on exploratory text mining in support of open-ended analysis, which fits the aim specified by the research question. Moving into uncharted text mining territory, the challenge presents itself in a lack of knowledge resources, and often also in a lack of knowledge about the analysis objective. Targeted analysis, such as prediction, requires sufficient understanding of the problem and data, which, as Tukey [202] asserts, exploratory analysis can provide. I further argue that the qualitative and information-rich nature of text motivates a general exploratory approach to text analysis. Considering the similarity in setting to knowledge-lean problem solving, in psychological terms,[4] it seems fitting to describe this form of text mining as knowledge-lean. Although standing on the shoulders of giants, this is text mining *from scratch,* in a knowledge intensiveness sense.

1.3 Overview of the thesis

This chapter has so far introduced the focus of the thesis in general terms. The next chapter delineates theoretical foundations from three main fields of science that support the discussion in the third and main chapter. Chapter 3 presents different types of analysis, the computational methods that support them, and ties the methods together by the characteristics they share. Chapter 4 then describes three different areas of application for these methods, and connects the methodological discussion with the work described in the included publications. Finally, I reflect on the work, limitations and future directions.

1.3.1 Disposition of publications

The chapters present the work which has originally been published in the form of the six papers included in the second part of this thesis.[8] The bibliographical information of these publications is also listed in the beginning of the book. How these papers relate to one another is explained and illustrated below.

Figure 1.2 provides an overview by mapping the individual Papers I-VI in a diagram. As this work is presented in Chapters 3 and 4, which describe a methodological and an application-oriented perspective respectively, the figure illustrates these two dimensions accordingly. It also illustrates the chronology of the work and the methodological progression, with solid arrows indicating explicitly which papers share methodology and dashed arrows which provide extensions motivated from the application perspective.

[8] All papers are reprinted with permission by their respective publishers.

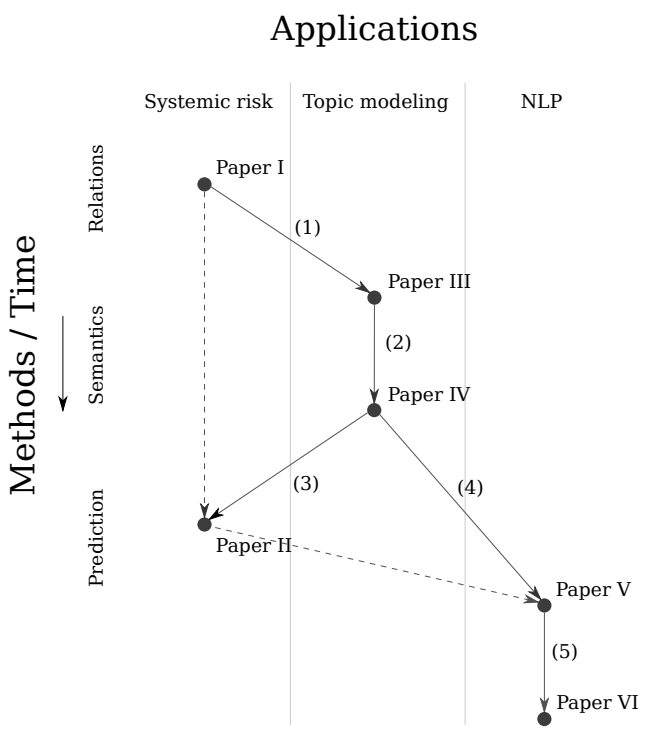


Figure 1.2: Overview of the papers of the thesis. The papers (nodes) are organized vertically according to methodological focus and chronology (when the work was initiated), and organized horizontally according to area of application. Solid arrows show methodological dependencies and dashed arrows show dependencies that relate to the application.

Paper I (bank networks) applies relation modeling and network visualization (Section 3.2) in analyzing systemic financial risk. Paper III (topic model visualization) extends upon Paper I by introducing semantic modeling and combining it with network visualization. Paper IV (topic modeling with word embeddings) extends upon Paper III by introducing word-level semantic modeling. Paper II (event detection and description) introduces predictive modeling based on semantic modeling similar to that of Paper IV, and extends upon Paper I in terms of application. Paper V (feed-forward discourse parsing) also performs predictive modeling and the semantic modeling methodology of Paper IV, applied to natural language processing and the task of discourse parsing, which is relevant as a venue to refine results from Paper II. Paper VI (recurrent discourse parsing) extends Paper V by introducing recurrent modeling.

Paper I is an extension of a previous paper,[9] listed as *other co-authored publications* nr. 4, and of a poster [177]. The full manuscript is also published as European Central Bank Working Paper No 1876.[10] Paper II is an extension of papers nr. 1 and 2.

1.3.2 Summary of contributions

As this thesis is based on co-authored publications, my personal contributions to these works are specified in the following. I have had a leading role in the conceptualization and realization of all my first-author papers.

In particular, in Paper I, I am responsible for data preparation, methodology and modeling (excluding that based on accounting data), as well as for all graphics and writing substantial parts of the paper. In Paper II, I am responsible for data preparation, method, modeling, graphics, and writing a large majority of the paper. In Paper III, my co-authors contributed in planning of the application and its evaluation (reported in an unpublished extended version and summarized in this thesis). I am solely responsible for Paper IV. In Paper V, I am responsible for the neural-network-based methodology and modeling, as well as writing the related parts of the paper primarily. Finally, in Paper VI, I am responsible for the method and modeling (with some support from co-authors and others), parts of the graphics, and writing substantial parts of the paper.

[9] This version was awarded the 2014 IEEE Computational Intelligence for Financial Engineering and Economics (CIFEr) 1st Best Student Paper Award.

[10] https://www.ecb.europa.eu/pub/research/authors/profiles/samuel-roennqvist.en.html

Chapter 2

Foundations

"Language shapes the way we think,
and determines what we can think about"

– Benjamin Lee Whorf (1897-1941)[1]

Representation of meaning plays a very central role in this thesis, be it in the form of human language and its expression in text, in internal representation in machine learning, or in visual form for communication. All these may be considered different forms of language, as they provide means for communication and reasoning. Form of representation has a significant impact on how information can be expressed, processed and understood, and understanding the nature of representations is fundamental for their effective use.

Before turning the discussion toward the main focus on text mining methods, this chapter presents some theoretical foundations from relevant fields, namely from linguistics, machine learning and information visualization. The chapter starts by reflecting upon the nature of human language and how it may be approached, as a basis for the development of more sophisticated text mining tools. Then, introductions to machine learning and visualization describe further fundamental concepts that underpin the work presented throughout the thesis.

2.1 Linguistics

Approaching text from the perspective of data mining, as discussed in Section 1.1.2, it is tempting to opt for the most straightforward computational

[1]The quote is widely attributed to Whorf, and writing about his work, Stuart Chase paraphrases it in [216].

methods able to extract information of interest, and disregard a linguistic perspective on text in the process. Nevertheless, viewing language through a linguistic lens is rather different from the highly pragmatic and quantitative inclination of text mining. Shifting the perspective may be beneficial, offering guidance going forward as linguistically more naive approaches may be reaching their limits. Computational linguistics incorporates a linguistic understanding as it assumes a rather pragmatic approach in building tools and resources for natural language processing. These tools then serve to make text mining systems more sophisticated and language aware. Not only does understanding language as a phenomenon support work on natural language processing, but it may offer healthy reflection for text mining design as well, by offering a more holistic view.

2.1.1 The nature of language

Widdowson [91] writes, "the essential nature of language is cognitive", and that it is a cognitive construct for abstract knowledge representation internally, as well as a means for communication and interaction with the external world. Language is a classification system of the world that we experience individually and collectively. It provides abstract categories and organization for concrete experiences, which Widdowson calls *conceptual projection*, a means for people to "cope with the third person reality of events and entities" by reference to the shared classification system that language provides. Thus, language simultaneously holds a personal/cognitive and interpersonal/communicative function. This function requires language to be inherently dynamic and flexible.

Two important design features that provide language with such flexibility are, according to Widdowson, *arbitrariness* and *duality*. Arbitrariness means that linguistic form does not resemble its meaning, rather meaning is arbitrarily assigned and linguistic units therefore can map freely to the abstract concepts of language. Meanwhile, duality provides that meaningless elements combine to form meaningful units, e.g., letters to words, which is a source of infinite productive power. Together these two features provide unlimited potential for generating references to abstract concepts as need arises. This flexibility of naming concepts, and freely composing expressions with them, is what brings the tremendous variability of language about. The prevalent ambiguity and implicitness of language constitute further challenges to its systematic analysis.

2.1.2 Linguistic analysis of text

Linguistics, as most fields of science, is concerned with constructing models from observations in order to understand and describe phenomena. This

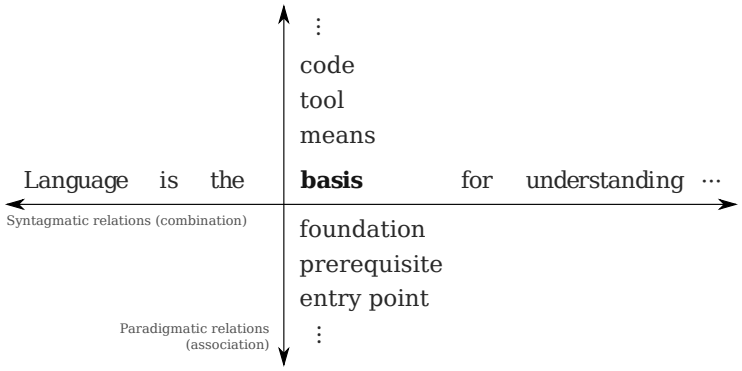


Figure 2.1: Dimensions of linguistic analysis: syntagmatic combination (horizontal) and paradigmatic association (vertical) of linguistic units (e.g., words).

creates a separation between the idealized model and real, often messy observations. In linguistics, there is an idea of an invariant form of a language that its speakers share, and a focus on characterizing these regularities of language rather than individual or incidental particularities (cf. de Saussure [59] *langue* vs. *parole* and Chomsky [52] *competence* vs. *performance*).

Linguistics deals in abstract *types* whose instances are *tokens*, Widdowson [91] points out, and it is concerned with classifying and organizing these types, based on some fitting measures of similarity and grounded in the observation of language use. Linguistic analysis can be employed at many different levels, and at each level there are two kinds of relationship among types, i.e., two dimensions of analysis [59].

Dimensions of analysis. Horizontally, linguistic elements combine in a *syntagmatic* relationship, e.g., there is a strong tendency governing what word may follow a given sequence of words. Vertically, elements associate in a *paradigmatic* relationship, e.g., different words may have similar likelihood of appearing in a given context, and are in a sense interchangeable. The two dimensions are illustrated by an example in Figure 2.1, showing relations for a single word.

Types form constituents, not only at word level. For instance, phrase structures form syntactic classes above the word level, where different phrases of the same class are paradigmatically associated along the vertical dimension. These phrases are then functionally equivalent (in their syntactic context), but may be structurally very different (e.g., consider replacing *basis* for *supposed basis* or *most useful tool* in Figure 2.1). The freedom along the vertical dimension allows for endless variation along the horizontal dimen-

sion. This way of freely and recursively linking units of meaning gives rise to the expressive power of language.

Levels of analysis. Linguistic analysis involves decoding from the linear surface form of language in order to construct meaning as a deeper structure. In analysis of text, meaningful linguistic units are found at the word level, as well as below and above. At the *morphological* level, linguists study how morphemes function as constituents of lexical units [136], e.g., how a word stem with a meaning of its own is modified by an affix. A word's meaning is further extended and modified at the *syntactic* level, as words are combined into phrases and sentences.[2] *Semantics* as the study of how meaning is encoded in language thus may operate at multiple levels. *Lexical semantics* analyzes the literal meaning of lexical units [58], e.g., words or compound words, and organizes them as well, by relations such as synonymy and hyponymy. Words may drastically change their meaning as compounds, but also adjust in meaning in syntagmatic combination, as they join higher-level structures. *Componential analysis* seeks to disentangle features of meaning of lexical items, visible as in compounds or hidden, whereas *compositional semantics* deals with how words determine and assemble meaning in the context of a sentence. [135, pp. 110-111]

Beyond the focus of semantics on what language means, *pragmatics* concentrates on what people mean by language. [124] This moves into the realm of language as a social instrument, acknowledging that meaning cannot always be fully understood without its social context, where language is used to serve a purpose. At his level, Widdowson [91] writes, analysis approaches interpretation as it moves from studying the text to its context, in trying to infer the *discourse* that the text is part of.

This level of analysis approaches the limits of linguistics and what may be learned easily from linguistic data. As Widdowson states, "the greater the units we deal with, the less we idealize the data and the closer we get to the actuality of people's experience of language", this implies the challenge of systematically and computationally analyzing discourse compared to lower linguistic levels, as it requires more human-like capabilities and may become more subjective in the process.

2.1.3 Computational linguistics and natural language processing

Computational linguistics studies and constructs computational tools for language analysis and generation, as a way of experimenting with and understanding language phenomena, as well as to produce practical tools [82].

[2] Nonetheless, linguistic inquiry into syntax is often focused on formal aspects of grammar in isolation from word meaning (cf. [52, 164]).

It entails a more applied focus compared to the general linguistic perspective on language. Natural language processing (NLP) may be considered the engineering aspect of computational linguistics, or a field in its own right. The two are closely related, and natural language processing draws upon computer science and machine learning to a large extent in its pragmatic focus to produce and improve natural language processing tools.

Natural language processing has been applied extensively at all the levels of linguistic analysis discussed above, and has been especially successful at the lower levels that favor idealization and empirical quantification. Common natural language processing software perform syntactic analysis such as part-of-speech tagging, syntactic and dependency parsing [114, 47, 5], as well as morphological analysis [34] and semantic role labeling [56, 29]. Text processing systems may also exploit software and database resources for lexical-semantic analysis [145], sentiment analysis [40, 191] and coreference resolution [54], to name a few. Such tools are usually developed using annotated text corpora and other linguistic resources, which require substantial manual efforts by experts. Therefore, it is difficult to adapt tools to new languages or even new domains or text genres, e.g., from newswire text to social media text, scientific articles or historical documents. Nevertheless, these kinds of tools serve as valuable components in many text processing pipelines, including for text mining purposes, and generally represent a bottom-up approach to language analysis that helps structure text data for practical applications.

At the discourse level, computational linguists also assume what might be considered a bottom-up approach (guided more by the text than its context) compared to the (socio)linguistic view (cf., e.g., [205]), by studying how sentences relate among each other. Thus, computational discourse analysis operates closely above syntax and often across sentences. It may serve to disambiguate individual sentences, as well as to link information across sentences, to form a representation of the discourse of a text. This may take the form of resolution of anaphoric references, parsing of discourse structure, etc. [108, pp. 671-707] While various linguistic frameworks for representing discourse structure have been proposed that pursue deep data structures or strict logical formalisms [133, 120, 213], more applied efforts have focused on *shallow discourse parsing* for recognizing relations between sentences (see [171, 228, 221] and Section 3.4.2).

Yet, computational discourse analysis in separation from the context of social code and real-world knowledge may eventually prove difficult. Grishman [82] points out that connections between sentences usually are implicit in the text, and that this creates ambiguities that require background knowledge in order to be resolved. The representation and utilization of such background knowledge constitutes a challenge to discourse analysis, which eventually may call for more general forms of machine intelligence.

Understanding text and its structure at a more global level may be helpful especially for exploratory applications of text mining. Discourse analysis does not necessarily need to rely on a bottom-up pipeline,[3] which might become unreliable and infeasible due to the complexity inherent in longer pipelines, but rather be approached in a top-down fashion. Topic modeling and discourse parsing are two ways of organizing text at a higher level of abstraction, both of which are addressed in this thesis (see Sections 3.3.1/4.2 and 3.4.2/4.3). Whereas topic modeling organizes text in terms of broader themes in a fully data-driven manner, discourse parsing may serve to link such themes and help meaningful representation by contextualizing themes and expressions in a discourse structure (e.g., for text summarization [95]). Such combination illustrates how linguistic understanding and natural language processing can enrich otherwise linguistically agnostic text mining approaches.

2.2 Machine learning

The idea of machine learning is, rather than explicitly programming computers, to let them adapt some of their behavior automatically by learning from data and experience, i.e., it is based on inductive inference [157, p. 398]. Machine learning has its early roots in statistics and probability theory (e.g., Bayes' theorem [15] in the 18^{th} century and least squares method for linear regression by Legendre [123] and Gauss [76] in the 19^{th}), as well as in early work on pattern recognition for engineering applications in the 1950-60s (cf. [157, p. 62]). In artificial intelligence, machine learning represents a focus on data-driven methods in contrast to the symbolic, logic-based paradigm, i.e., it often relates to perception tasks rather than simulation of higher-level cognitive tasks. As such, machine learning has gradually grown more central to the field of artificial intelligence [157]. In data mining, machine learning is a vital component for performing modeling and analysis of data (cf., e.g., [223]). It is equally important to text mining, where it serves natural language processing in handling uncertainties and irregularities of language patterns, as well as higher-level pattern recognition and data mining operations.

Machine learning is applied in various ways for different types of tasks and data. These types of learning and some fundamental concepts are outlined in the following. Machine learning encompasses a range of learning algorithms and families thereof. I continue the discussion focusing on the family of neural networks, which is of primary relevance to this thesis.

[3] For examples, confer the linguistically resource-lean systems participating in the CoNLL 2016 Shared Task [221], mentioned in Section 3.4.2.

2.2.1 Principles and types of learning

Supervised learning. The first major type of learning, called *supervised learning*, is relevant for modeling the relationship between a set of variables and a target variable, that is, a model is constructed to predict an output variable of choice based on a set of inputs. Depending on whether the output variable is quantitative or categorical, this type of modeling is referred to as *regression* or *classification*. In this thesis, I often refer to supervised learning as *predictive modeling*, in order to emphasize the process of estimating one variable based on others, irrespective of how they relate in time. By contrast, I refer to prediction forward in time as *forecasting*.

The learning process is supervised because, when feeding a model with input data, the output can be compared against the output variable in the data set. The discrepancy between the observed value and the estimated value constitutes a measurable error, which is defined through a *loss function* [200]. The goal of the learning process is then to optimize parameters of the model with respect to the loss function, over a *training set* of input-output pairs. Parameters are variables of a model that are fitted, or learned, such as the variables w (weights) in the linear combination f:

$$y' \approx y = f(x) = w_o + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

where y is the output variable and x a vector of inputs for one sample.

The equation defines a linear regression model (f) whose output is a quantitative variable. The loss function typically measures the difference between the predicted value y and the observed value y' in the data over several training samples, as the parameters are optimized by some training algorithm.

The model can be repurposed to serve as a classifier as well. As a simple example, if two classes are represented as $y \in \{0, 1\}$, we may classify samples as positive ($y = 1$) by a threshold $f(x) > 0.5$ (cf. [200, p. 101]). While the loss function guides the learning process of a classifier, more intuitive measures of error are based on the correspondence between class predictions and observations that can be summarized in a *confusion matrix*.

Most notably, *accuracy* is the rate of correct classifications into either class based on all predictions. The threshold on $f(x)$ determines the sensitivity of the classifier, which implies a trade-off between completeness and exactness of positive classification, measured as *recall* and *precision* rates respectively. The setting in which the classifier is applied may imply a preference toward either completeness or exactness. For instance, a supervisory setting typically may call for higher sensitivity, while accepting a higher rate of false positives.

After training, the predictive performance of the model may be assessed on a *test set* of the same form as the training set, but separate from it. This

provides an estimate of how well the model can be expected to perform on previously unseen data, for which the correct output is unknown. In addition, a *validation set* may be used for *model selection*, namely to find and pick the best model when several are available [200, p. 222]. The model presented above assumes a fixed linear form and is fitted only by its parameters. Other types of models often allow for a higher degree of flexibility (or model complexity), which is defined by one or more *hyperparameters*. Model selection implies setting hyperparameters in order to achieve an optimal balance between flexibility and ability to generalize, as measured by model performance on a validation set. This procedure controls for *overfitting,* which occurs when a model has so much freedom to adapt to the training data that its ability to represent general patterns and predict on unseen data is deteriorated (cf. [200, p. 219]).

Unsupervised learning. The second major type of learning is called *unsupervised learning* and is used when no suitable target variable is present, and only the input variables are used for modeling. While supervised learning supports predictive modeling and targeted analysis, unsupervised learning serves exploratory analysis as well as it serves as preprocessing for supervised learning.

Clustering, a common form of unsupervised learning, is the task of finding natural groupings in data and assigning clusters, based on some measure of similarity among data points. It is analogous to classification in aiming to simplify data by categorizing it, but unlike labeled classes, clusters do not come with explicit interpretations.

Without a designated target variable, all variables are treated equally, as dimensions of a space where similarity can be defined as distances (e.g., Euclidean distance or vector angle cosine). Similar to the loss function for supervised learning, a scoring function is defined that generally aims to minimize distances within clusters and maximize distances between clusters. Some clustering algorithms assign crisp clusters, i.e., each data point belongs to only one cluster, whereas others perform fuzzy clustering, where cluster membership may be represented as a probability $p(c = i|d)$ for each data point d and cluster $c \in [1, k]$. The number of clusters k is often given as a hyperparameter, whereas some algorithms seek to infer an appropriate number of clusters, or represent their clustering as hierarchies. Confer Zaki & Meira [223, p. 28] for an in-depth review of clustering paradigms and algorithms.

Another form of unsupervised learning is *dimensionality reduction*, which similarly aims to simplify data. While clustering reduces from data points to a lower number of clusters, reduction of dimensionality means representing high-dimensional data in a lower number of dimensions while minimizing

distortions and loss of information.[4] It may be used for visualization by projecting data into one, two or three dimensions for plotting, as well as in conjunction with other types of modeling, where it may serve as a means of compacting sparse data. For further reading confer [223, p. 183] and [200, pp. 528-575].

Unsupervised learning is in fact performed by supervised learning methods at times, when the same type of data is used both as input and as output labels. For instance, in *language modeling*, classifiers are trained on sequences of tokens to predict the next [19] (a way of modeling probability distributions of sequences through self-supervision), or autoencoder neural networks are used to reconstruct its input through prediction using lower-dimensional intermediate representations [94]. Both approaches perform dimensionality reduction, as they produce dense low-dimensional representations of sparse high-dimensional input (e.g., words in a text).

Learning representations of symbolic language data is the essence of how machine learning is employed in this thesis. It provides a data-driven way of modeling semantics and dealing with the challenge of symbolic data, that is to say, it offers a means of comparing and organizing symbols that are otherwise unrecognized and all distinct from one another.

2.2.2 Neural networks

The name *(artificial) neural network* derives from the biological inspiration for a family of models for information processing based on conceptual versions of neurons and neural connectivity. Theorized by McCulloch & Pitts [137] in the 1940s, the first artificial neural networks were implemented in the 50s by Minsky [148, 149] and Rosenblatt [178]. These networks, which at first were implemented using analog hardware and not the digital computer, established a paradigm for neural information processing in parallel to logic-based computing. Rosenblatt's *perceptron*, as he called the model, can be described as a linear combination of inputs with a non-linear activation function (a step function) [27, p. 192]. By extending the linear model described above with an *activation function* σ, the perceptron can be generalized as:

$$f(x) = \sigma(w_o + w_1x_1 + w_2x_2 + \cdots + w_nx_n)$$

With a logistic activation function, this formulation becomes equivalent to *logistic regression,* set forth by Cox [57] in the statistics community in 1958. The (single-layer) perceptron met a lot of enthusiasm in the artificial

[4] Some methods perform both data and dimensionality reduction, such as the Self-Organizing Map [115].

intelligence community during the 1960s, but was then increasingly criticized toward the end of the decade, with Minsky & Papert [150] famously demonstrating its inability to handle problems that are not linearly separable (such as the XOR function). Subsequently, the focus of the field largely shifted in favor of symbolic approaches to artificial intelligence.

In the following years, procedures for model fitting by back-propagating errors were being proposed by several researchers,[5] and Rumelhart et al. [179] were in 1986 able to popularize the *backpropagation* algorithm by showing that it can be used to effectively train multi-layer perceptrons. Using a differentiable activation function (e.g., logistic) and the chain rule, back-propagation computes error gradients with respect to weights throughout a network, across multiple layers. The error gradients enable optimization of the weights by an optimization algorithm, such as *stochastic gradient descent*. Because multi-layer networks are able to approximate any function, the limitations of the single-layer perceptron could be overcome [27, p. 230][200, p. 390]. Unlike many other machine learning methods, including linear models with basis expansions and Bayesian methods, neural networks do not require prior assumptions about the distribution of data to be made, which make them more flexible but also more data intensive.

The multi-layered *feed-forward neural network* consists of a simple *input layer*, typically one *hidden layer* of neurons, and an *output layer* of neurons. Hence, the perceptron/logistic regression model may be duplicated within the output layer to create multiple output signals, and each output unit is fully connected to all units in the previous layer, which likewise connect fully with the input units. When two layers are fully (or densely) connected, each node in one is connected to all nodes in the other. For simplicity, this pattern of connectivity is often illustrated by a single arrow between the layers, which are drawn as blocks that hide the individual nodes. The three-layer network can be formulated as:

$$\mathbf{y} = \sigma^{(2)}\left(\beta^{(2)} + \mathbf{U}^{(2)}\sigma^{(1)}\left(\beta^{(1)} + \mathbf{U}^{(1)}\mathbf{x}\right)\right)$$

with input vector $\mathbf{x}$, and trainable weight matrices $\mathbf{U}$ and biases β as parameters. The logistic activation function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Letting each output unit represent a class, the neural network offers a straightforward approach to multi-class classification. Setting the outer activation function to the *softmax* function,

[5] Linnainmaa [127] introduced the method in a general form by 1970, and Werbos [215] applied it in the context of neural networks, inspired by Freud's concept of backward flow of credit assignment. Others followed on this path, including Rumelhart et al. [179].

$$\sigma^{(2)}(x)_i = \frac{e^{x_i}}{\sum_{c=1}^{C} e^{x_c}}$$

transforms the output to a probability distribution that provides an interpretable basis for decision confidence. The standard classifier then chooses the most probable class from the distribution.

2.2.3 Deep learning and representation learning

While neural networks with one hidden layer in theory are able to approximate any function, this may in practice require infeasibly many hidden units and large amounts of data. Instead, machine learning has been widely supported by the practice of *feature engineering*, whereby expert knowledge is utilized to transform and select relevant parts of the input before it is fed to a machine learning model, in order to reduce its variability and thus the amount of data required for estimation [122]. Bengio et al. expand:

> "The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied. ... [F]eature engineering is important but labor-intensive and highlights the weakness of current learning algorithms: their inability to extract and organize the discriminative information from the data. Feature engineering is a way to take advantage of human ingenuity and prior knowledge to compensate for that weakness." [18]

Nevertheless, the question remains whether, instead of relying on human learning and intelligence, this part of the process can be automated and machine learned, too. They continue:

> "An AI must fundamentally understand the world around us, and we argue that this can only be achieved if it can learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data." [18]

As an alternative to designed representations, the *representation learning* approach that Bengio et al. [18] advocate seeks to learn appropriate representations automatically, such that features that are relevant for the classification problem are highlighted, whereas irrelevant variations are suppressed for subsequent learning steps. This idea is the essence of what is called *deep learning*, an umbrella term whose name refers to its typical implementation using deep neural network architectures [184]. Such networks may have multiple hidden layers in a feed-forward arrangement that can learn hierarchical representations that compose layer by layer.

LeCun et al. [122] write that training deep networks by backpropagation at first turned out to be unsuccessful in most practical cases, as the procedure had difficulties in effectively optimizing layers further from the output (due to the vanishing gradient problem). Interest was however revived around 2006, they describe, due to successful use of unsupervised learning for pre-training of deeper layers, which along with increases in computational power and data availability has sparked many efforts to train deep networks since, resulting in breakthroughs in areas such as computer vision [117], speech recognition [92], machine translation [195] and other applications of natural language processing (e.g., [56, 190]).

They point to a particular strength of representation learning being its ability to generalize well with limited labeled data, as it uses unsupervised learning on more extensive unlabeled data to capture meaningful regularities, by means of vectors holding *distributed representations*. The distributed nature of representation plays a key role for generalizability. Latent aspects of the data are learned and represented as a vector of their strengths, effectively enabling comparisons of symbolic input (as first explored by Hinton [93]). Natural phenomena with a componential structure benefit from the combinatorial power among these latent dimensions, and deep layer structures enable hierarchical composition and exponentially more expressiveness of distributed representations [17, 18]. Such compositional and componential structures are abundant in natural language, as discussed in Section 2.1, at various levels including sentence composition and non-compositional semantic structures of words.

In a neural setting, language is typically modeled by predicting words from their context of previous words, using an internal distributed representation of words (cf., e.g., [19, 143, 142]). The representation, generally called a *word vector* or *word embedding* (see Section 3.3.2), provides an embedding into a semantic space, where words can be meaningfully related. The concept of an embedding can be extended to other types of entities such as sequences of words [121] (see Section 3.3.3), word senses [173], users [20] and other types of categories. Although embeddings are commonly distributed and provided by neural architectures, other noteworthy approaches have been explored as well (for an overview and comparison see [13], and examples [26, 163]).

In modeling language and other sequential data, *recurrent neural networks* have proven especially helpful as they offer a practical way of integrating information in a sequence at arbitrary distances. As illustrated in Figure 2.2, recurrence is implemented by connecting a hidden layer to itself, such that its state h_t after receiving input x_t at a time step t is fed back as input in $t + 1$, along with new input x_{t+1} from the previous layer. The hidden layer thus holds a continuous and distributed state representation h, which also feeds to the following layer in order to inform the final output.

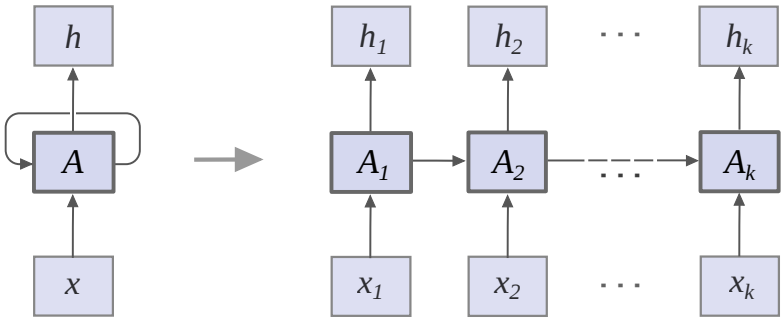


Figure 2.2: Illustration of a recurrent neural network layer (A), to the right unrolled over the input sequence of vectors x. The arrows indicate fully connected layers/time steps.

Backpropagation is used to train the network from output to input, where depth equals the number of past time steps and is arbitrary [184].

In practice, standard recurrent neural networks however have proved unable to learn long-range dependencies. The problem was first solved by the extension of Long Short-Term Memory (LSTM) [96] to recurrent networks, which provides a memory cell controlled by gates for reading, writing and forgetting. Gate activation is itself learned by backpropagation and allows selective access and modification of the distributed representation of the cell, effectively allowing the LSTM network to model dependencies over hundreds of time steps, which makes them highly practical for natural language processing [122, 18, 110].

More recently, further extensions to deep neural architectures have helped expand the capabilities of machine learning, even beyond its typical kinds of uses of classification, regression, clustering, etc. These extensions include attention mechanisms (e.g., for describing relevant parts of an image in natural language [220]) and more sophisticated memory modules that allow for inference of simple algorithms (e.g., the Neural Turing Machine [78]) and reasoning based on more complex data structures (e.g., Differentiable Neural Computer [79] for navigating graphs, and AlphaGo [187] surpassing human performance in the game of Go). Advances like these are widening the application of deep neural networks from perception tasks toward mimicking higher-level cognitive functions, and as neural computation is being applied more generally it is entering domains traditionally ruled by logic-based computing. Also for data and text mining, this trend points toward machine learning increasingly being employed for data-driven, adaptive reasoning, in addition to pattern recognition.

2.3 Information visualization

Visualization constitutes a cornerstone of this thesis, as a means of communication between computer and mind (a visual language), and as a basis for human-computer cooperation. While machine learning provides meaningful abstraction of data, visualization serves as an interface toward the human, in order to integrate her into the data analysis process. This section presents basic principles of *information visualization* relating to human visual perception and cognition, and visualization design, including the role of interactivity. This leads to discussion of the integration of data modeling within the scope of *visual analytics*, which provides this thesis a general framework for combining machine learning and human reasoning for data analysis (detailed in Section 3.1).

Information visualization as a field aims at supporting people in making sense of data by means of visualization, by mapping data to visual representations that help us understand its structures and quantities [211]. It is the study of how to design visualizations to best fit human perception, cognition and search of meaning in data, which requires a theoretical and practical understanding of how people see and reason based on visuals. Information visualization generally concerns visualization of abstract, statistical data, requiring design choices to be made regarding the meaningful mapping of different variables to visual representations. The choice of such *visual encodings* is guided by principles proposed within the field, which are based on theoretical understanding of the human visual system, while more practical observations offer guidelines in particular for the design of the visual interface as a whole [153].

2.3.1 Visualization and perception

The human visual system processes visual information from the eyes in multiple stages, from the recognition of low-level features toward higher-level structures of meaning, in order to integrate it with other functions of the brain. As the different levels of processing target different levels of complexity in the information, and at different speeds, designing effective visualizations implies matching visual encodings and layout with the capabilities of the visual system. At a broad level, encodings can be divided into *sensory* and *arbitrary symbols* [210, p. 10]. On the one hand, the human visual system has through evolution become highly specialized at perceiving certain low-level features as meaningful sensory symbols. On the other hand, we have through our everyday experiences learned to understand other arbitrary symbols, which are meaningful only within certain contexts or cultures. The former are universal and fast to process, while the latter are expressive and capable of change (compare with arbitrariness of language, Section 2.1.1).

Sensory representation offers a limited scope for the encoding of information, but should be used whenever possible, as it enables high-speed and parallel recognition of elementary features in the visual input stream[6] by the primary visual cortex (V1).

Feature recognition. Visual processing at this early stage is called *preattentive processing*, as it is done effortlessly and quickly without focused attention [199]. At this stage, the visual cortex recognizes simple and local features such as size, orientation, motion and color over all receptors in the retina in parallel. By these features, areas in the field of vision that stand out can be detected and attention directed, in order to initiate the serial mental processing following this stage. The local-feature processing mechanism influences how we perceive different types of visual elements, in concord with subsequent steps that focus on larger visual units.

Pattern recognition. The perception of simple features relates to the concept of *visual variables* [21], also called *visual channels* [153], in the visualization literature. This theory describes what types of visual encoding are suitable for what type of information and purpose. For instance, color hue or shape is useful to associate objects to each other or make some objects stand out against others, whereas it is easy to compare quantitative differences in length or volume.[7] The most important visual channels include spatial position, color hue, intensity, size/length/width and orientation.

Moreover, visual representations are constructed using marks (e.g., points, lines, areas) that represent entities, relations or sets. Information is made visible by marks and their alterations in a visual channel. For example, a point can encode values by its position, size or color, whereas a line is suitable to encode a relationship and can encode a related value by width or color.

Visual working memory. Following the feature recognition stage, the visual cortex performs pattern perception. It is a dual process where features are assembled into simple patterns and compound shapes bottom-up,

[6]Sensory signals come primarily from the center of the retina, the fovea, but also from more sparse receptor areas in its periphery. The eyes move across the field of view in patterns of quick saccade movements (ca 20-180 ms) and short periods of fixations (ca 200-400 ms) in order to sample key points of the scene in high resolution and construct the impression of a complete scene that we have. The movements are subconsciously controlled and during a saccade vision is suppressed, which means that details or changes occurring outside of the point of fixation or during a saccade can easily go unnoticed [210, p. 141]

[7]Munzner [153, p. 99] identifies channels of two types: identity channels for categorical information, and magnitude channels for quantitative information.

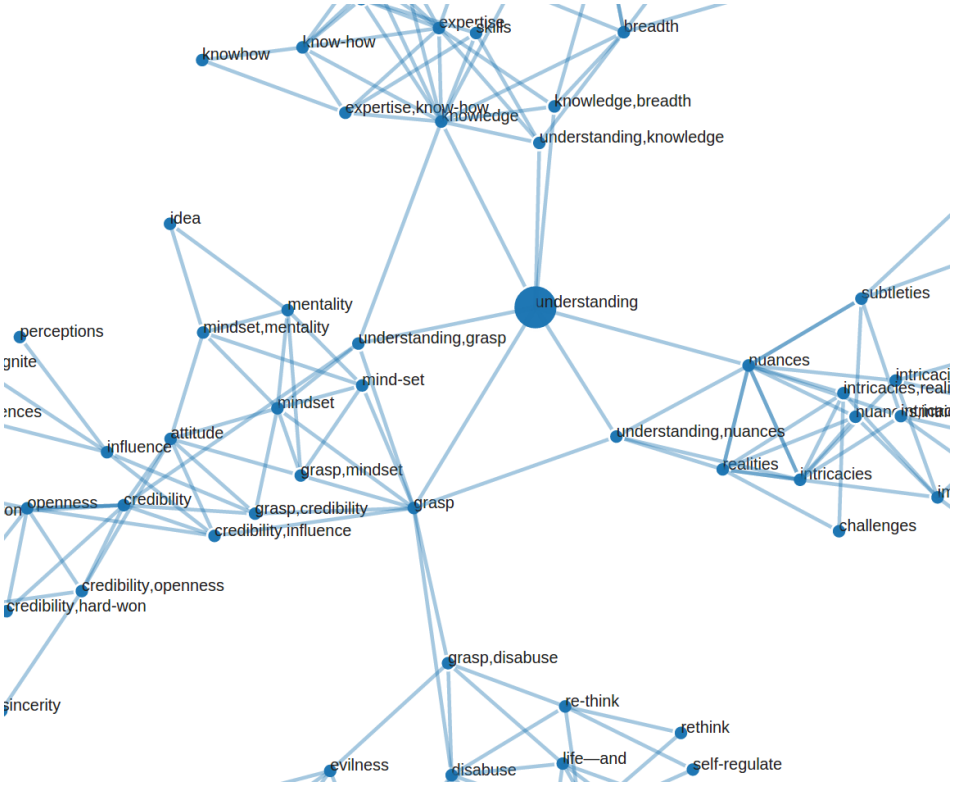


Figure 2.3: Example visualization using the principles of spatialization and connectedness to illustrate similarity structures close to the word *understanding* in a high-dimensional semantic space.

while attention reinforces relevant information top-down [210, p. 21]. Immediately after the visual pathway, the inferior temporal cortex contains specialized neurons recognizing specific learned patterns such as letters, faces and familiar objects. Spatial information is decoded by a parallel pathway that connects to centers for motor control.[8] Our highly developed sense of space is often exploited to convey similarity structures in data. Using spacial distance as a visual metaphor for similarity is referred to as *spatialization* [165] and can be used to effectively communicate structures even in non-spatial data, e.g., in conjunction with dimensionality reduction techniques (see example in Figure 2.3). Under the name of *Gestalt laws* [210], further principles have been identified that specify various types of patterns that are easy for humans to isolate, such as connectedness (e.g., node-link diagram for network visualization; see Figure 2.3 and Section 3.2.3), similarity

[8]This high degree of specialization towards interpreting spatial information and acting upon it (e.g., by eye or hand movement) may be contributing to the fact that the spatial dimensions generally are the most powerful means of communication in visualization, as compared to other visual channels that are preattentively recognized but the information of which might be understood only at higher levels.

(e.g, in shape), continuity (e.g., connected lines, vector fields), symmetry and closure (e.g, Venn and Euler diagrams for visualization of sets).

Information visualization provides a framework for mapping data and information to visual representations in a way that strives for efficient decoding that supports analytical thinking, and that avoids distortion or misguiding representation (cf. Tufte's guidelines of *graphical excellence* and *integrity* [201]). Nevertheless, human visual perception involves imperfections that may lead to inconsistencies and misinterpretation. For instance, individual or context-dependent irregularities in color perception, contrast effects, spatial optical illusions, etc. may distort encoded information [210]. The effectiveness of visual representations are guarded by low-level principles, higher-level rules of thumb and eventually evaluation of implementations, e.g., through user studies [153, p. 67, 117].

2.3.2 Interaction, cognition and information seeking

Visual perception is to a large extent driven by sensory input, but it also relies on feedback from higher-level cognitive processes, in order to make sense of a scene by iteratively directing attention and eye movement, as well as by interacting and manipulating the visual scene. Following the pattern perception stage, the visual working memory attempts to associate various types of information with the incoming visual information in order to integrate it with other areas and already held knowledge[9] [210, p. 22]. It is done in a continuous and temporary fashion, constrained by a very limited storage capacity.

The visual working memory constructs visual queries that result in visual search strategies, directing attention in order to understand the visual [211, 210]. Interaction further facilitates understanding by allowing for manipulations of the view, such as highlighting of associated elements, navigation in a large space or reordering of objects. The visual interface is not only a medium of presentation, but also functions as an external memory augmenting our cognition and memory, especially working memory, where interactivity enables the necessary two-way communication [153, p. 6].

As one inspects, interacts and makes sense of a visual, observations are being integrated with held knowledge and a *mental model* is being constructed based on the visualized information. Increasingly, the elements of the visual assume the role of references to parts of the mental model. The visual display in front of the user contains details that can remind them about previous thoughts, partial insights or pieces of knowledge about the

[9]For instance, language information is processed separately from spatial recognition and object identification. These different types of information can be linked in visualization to help the user in understanding details quickly, e.g., by using symbols, text tool-tips or labels describing a data point or other element.

subject of the visualization, and iteratively support new thoughts and deeper understanding to arise.

Interactive visualization is meant to support *information seeking* (cf. [217]), as exemplified by Shneiderman's *information-seeking mantra*, "overview first, zoom and filter, then details-on-demand" [186], that crystallizes the idea of interactive exploration of vast data going from broad overview to details of interest. This principle is broadly applied to visual interactive interface design of today. Furthermore, while zooming, filtering or highlighting to focus on a specific detail, the surrounding information context is often kept at least partially in view, in order to create a smooth navigation experience that supports the users orientation in the visual information space, in effect combining overview and detail view simultaneously. Referred to as *focus-plus-context* [43], it is another example of design principles intended to support cognitive processes through visual display and its interactive modification.

Further principles state that interface design should direct attention toward the data or information rather than the interface itself, e.g., the surroundings should be kept consistent while visual marks are allowed to change [201]. Interaction should be responsive and non-distracting in order not to impede the visual thinking process, while supporting the cognitive thread and its series of nested feedback loops. Ware [211] writes about the economics of cognition, which relates not only to the idiosyncrasies of the visual system but also to the higher-level cognitive costs of learning to use new tools. Thus, visual interface design has to balance complexity in terms of visual detail and number of interaction possibilities against the amount of information one would want to convey. Interactivity drastically expands the space of information that can be conveyed on a limited screen. However, how to balance complexity and user freedom is dependent on the aim and setting of the visual interface, too, namely whether it serves more the purpose of narrowed-down communication or open-ended exploration.

2.3.3 Visual analytics

Visual analytics is an outgrowth of information visualization, and was originally proposed in 2004 as "the science of analytical reasoning facilitated by interactive human-machine interfaces" [218], making interactivity a defining property. Before that, the term *visual data mining* has been used to describe the extension of data mining with visualization [74, 112]. Much like the term analytics has become wide in scope (see Section 1.1.4), Keim et al. [111] observe that visual analytics has come to describe a multidisciplinary field spanning from data management to decision making, nevertheless, emphasizing the utilization of interactive visual interfaces and data modeling. Modeling in this practice includes basic statistics as well as machine learning.

Keim et al. [111, p. 10] define a visual analytics process for visual data exploration describing the relationships between data, models, visualization and knowledge, in terms of data mining, visual encoding of data and models and different types of user interaction. The process describing how to integrate data modeling and interactive visualization for exploratory analysis is further developed and discussed in Section 3.1, as a general conceptual framework for this thesis.

Chapter 3

Methods

"You shall know a word by the company it keeps"

– J.R. Firth (1890-1960) [71]

The computational methods discussed in this thesis rely on modeling of patterns naturally present in text, in particular the notion of co-occurrence plays an elementary role. The types of modeling range from the simpler to the more advanced, and from the more general to the more specific, in order to extract structure and meaning from text. Representation learning is a central theme, as is the integration of human analysis, in the pursuit of intelligent and scalable analysis of text that does not require extensive encoding of knowledge.

This chapter focuses on the commonality of the methods that underpin the work presented in the papers. It seeks to separate method from application, albeit a not always clear-cut distinction, in order to first emphasize and chart how knowledge-lean text mining can be approached, while more idiosyncratic details relating to a specific domain or setting are described in the next chapter. The discussion both reviews important previous work and describes extensions, which may be of general interest to anyone pursuing data-driven and knowledge-lean text analysis, as it explores different ways of addressing text by computational means in the context of human use.

3.1 A framework for joining computational and human analysis

The visual analytics approach, as introduced in Section 2.3.3, brings together human and computational capabilities for data analysis in a practical manner, and is conceptualized through the visual analytics process by Keim et

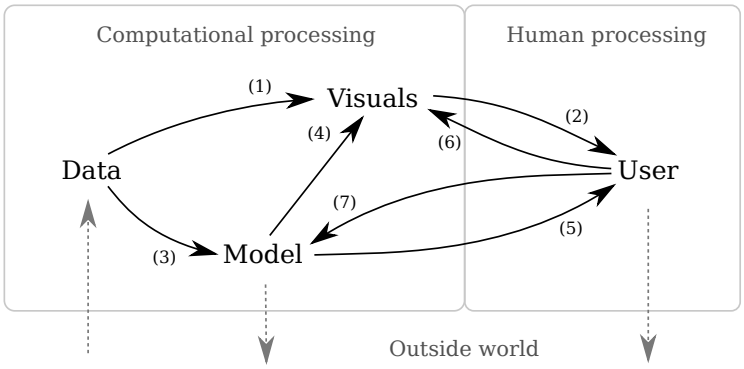


Figure 3.1: A conceptual framework for combining machine and human intelligence based on the visual analytics process.

al. [111]. In Figure 3.1, I show a reinterpretation of this process, which emphasizes and illustrates clearly the way computational processing interfaces with the human user, as parts of a joint data analysis process. The overarching goal of this conceptual framework is to describe how data can support the gaining of understanding. The framework rests on the concepts of *data*, *information* and *knowledge*. While these may have various interpretations (as studied within information science [229]), I choose to distinguish among them as follows. Data is information representing raw observations about the outside world, or treated as such. Information, as distinct from data, is any form of abstraction or idealization of data, e.g., a categorization or formal model. It is a representation that supports reasoning and communication. Knowledge is a human quality, an internalized form of information based on observation, which is not transferable other than through decoding (articulation)[1] into information. Understanding, with respect to new information, is then the integration of that information with held knowledge, or in other words, the formation of a mental model.

Reading Figure 3.1 from left to right, top first, data may be directly visualized by mapping to visual representations (1), as discussed in Section 2.3.1, and the visual representations are then shown to the user (2). When introducing modeling into the data analysis process, data is abstracted by constructing or training a formal model (3). The model can be presented to the user in non-visual forms directly (5), although emphasis here lies

[1]Nonaka & Takeuchi [158] distinguish between tacit and explicit knowledge, and refer to the transformation as externalization/internalization in their SECI model. Some kinds of knowledge are more easily externalized and communicated than others, for instance, Dreyfus & Dreyfus [65] distinguish fact-based *knowing-that* from intuition-based *knowing-how*, in their critique of expert systems and symbolic artificial intelligence that only focus on the former.

on fully exploiting the process by mapping information from the model to visual representations (4). Interaction at various levels introduces feedback channels into the process. Depending on the intended freedom of the user(s), interaction may target the presentation of information (6), model (hyper)parameters (7), data gathering and transformation, or the outside world. The most close-knit feedback loop is interaction with the visual for view manipulation (see Section 2.3.2). Parameter manipulation is also important for model-based data exploration, but typically comes with larger latency due to model training. Similarly, data transformation and other steps may all be subject to change as part of an iterative analysis and development process.

Finally, the process can be viewed from a decision-making point of view, in that the user may take actions based on acquired understanding, which affects the outside world. Models may also be set to autonomously act upon data (and internal state), although this is in stark contrast to the idea of visual analytics. Rather, the visual analytics perspective stresses the involvement of human experts and allows them to take charge of understanding the task and data at hand, and to perform a more nuanced, contextualized and versatile analysis than machines are generally capable of. Meanwhile, visualization as a high-capacity communication channel provides infrastructure for transparency of the model and data, which may be necessary, for instance, for accountability in decision making, and in general for building trust in otherwise more or less opaque models.

The framework serves to situate the methods presented in this thesis, and describe how they contribute to a broader analytics setting. The primary focus lies on the use of text mining methods to support text understanding within this context, while visual and interactive presentation has a supporting role. From a visual analytics perspective, modeling helps to structure and filter text data, as a basis for navigation and exploratory analysis. Natural language understanding is a particularly challenging task to automate, which therefore stands to benefit greatly from a cooperative approach between human and machine, especially when entering new text mining territory. Risch et al. [175] assume a similar approach to text exploration under the name of *visual text analytics.* The forms of modeling presented in this thesis stretch from network models to various machine learning models, and the use of visualization includes interactive interfaces as well as simpler static presentation. Developing and optimizing machine learning models is time consuming, as is the development and testing of interactive interfaces. The framework allows for a practical fluid balancing between computational and human responsibilities in the analysis process, as well as incremental development of the parts.

3.2 Relation modeling

Language can be viewed as fundamentally consisting of entities and relations among them. Entities as units of meaning are found at many different levels of linguistic analysis, as reviewed in Section 2.1. Whichever way we define the entities we deal with in a text, they stand both combinatorially in syntagmatic relation to one another, and associatively in paradigmatic relation to other entities in the language. Thus, modeling relations in text, one way or the other, is crucial to text mining. As text mining is interested in analyzing the content matter of text, rather than only modeling language, we should focus on connecting entities of the text to entities and other types of concepts of the world. Words, for instance, are interesting primarily as references to entities (e.g., nouns, proper nouns), actions (e.g., verbs, nominalized verbs) and qualities (e.g., adjectives, adverbs) of the world. At a very general level, text mining may be interpreted in terms of a relational model that defines certain types of entities, relations and attributes of interest.[2]

This section assumes such a general approach to relation modeling, centering around the data structure of a graph as the basic form of representation for different types of relations. Its use is first explored for the analysis of named entity co-occurrences, followed by the construction of network models based on those relations, and quantitative and visual analysis of networks. The main application of these methods is presented in Section 4.1.1/Paper I (construction of bank networks), while networks based on semantic modeling (Section 3.3) are used in the applications of Section 4.2/Papers III-IV (visual topic exploration).

3.2.1 Co-occurrence analysis

Co-occurrence analysis focuses on the syntagmatic dimension of text, and typically on the level of words, either for modeling of language or content. Within corpus and computational linguistics, frequent term co-occurrence patterns in language have been studied as *collocations*, i.e., idiomatic expressions such as *once upon a time*, semi-compositional expressions such as *give a speech*, fully compositional expressions such as *handsome man/beautiful woman* and compound nouns such as *black box* [69]. Within the natural language processing literature the terms *multi-word expression* and *n-gram*

[2]Many applications do not exploit all of these, e.g., extraction of semantic triples (subject-predicate-object) for construction of semantic networks [193] or knowledge graphs [138] may be cast as extracting entity pairs with predicate relations without additional attributes, or sentiment analysis [166] may be cast as attaching sentiment attributes to related entities (such as companies), without defining relations between entities. Meanwhile, for instance, biomedical event extraction [30, 4] generally defines events as relations (event types) with attributes (e.g., negation, speculation) among entities (e.g., genes and gene products).

Algorithm 1 Co-occurrence network construction (in: agg, csize; out: net)

```
for doc in agg:
  occs = ner(doc)
  for i in range(1, len(occs)):
    for j in range(0, i):
      if occs[i].entity != occs[j].entity and occs[i].index-occs[j].index < csize:
        net.links[occs[i].entity][occs[j].entity].weight += 1
```

among others are often used similarly. Evert [69] conducts a thorough study of statistical measures of term co-occurrence significance for modeling of collocations. He reserves the term collocation for a linguistically idealized form of the concept, while referring to the empirical forms as co-occurrences. Co-occurrences are often analyzed positionally, in terms of term adjacency, within longer spans, or within linguistic units such as sentences, paragraphs or documents. Alternatively, co-occurrences may also be analyzed relationally, e.g., based on syntactic relations. Co-occurrence frequency is generally normalized, by comparing against estimates based on individual term frequencies, in order to identify pairs that are significantly co-occurring.

For text mining purposes, co-occurrence analysis has been used to learn about the content matter of texts, as it provides informative relations among terms beyond the linguistic understanding of a co-occurrence or collocation. Although the methods are relatively simple, they have been key to early pioneering use of text mining in various fields. For instance, Wren et al. [219] analyze biomedical research articles and extract co-occurrence relations among entities such as genes, diseases, phenotypes and chemicals to support hypothesis generation, and Özgür et al. [162] identify co-occurrences among person names in news articles. Moringa et al. [152] study product reputation online based on co-occurrences between products and common phrases, and Magnusson et al. [131] similarly find co-occurrences between company names and other keywords to predict financial performance.

The method for co-occurrence analysis presented here focuses on modeling relations among named entities, i.e., names of things such as people, locations or organizations. In Section 4.1.1, the method is applied to study relations among banks in online discussions and news. Named entity recognition can be performed in various ways, e.g., based on generic features [70] or based on a dictionary of names. The method operates based on identified entities and a simple positional measure of co-occurrence. By not necessarily relying on natural language processing for the identification of linguistic units or relations, the co-occurrence extraction remains language independent and computationally very efficient.

The procedure is defined by Algorithm 1, which invokes a named entity recognition procedure (*ner*) that locates a sequence of occurrences in a doc-

ument (sequence of characters), in terms of entity identity and occurrence position (character offset). The algorithm receives as input the context size (*csize*) that defines a window in which co-occurrence relations are registered, and a set of documents, an aggregate (*agg*), containing the data. It returns a weighted network graph (*net*), which is initialized with zero weights between all node pairs. The corpus is partitioned into one or several aggregates and a network is formed for each. The aggregates may be organized, for instance, chronologically for dynamic network analysis. The network graph is a generic representation that supports the application of generic methods for network analysis, which is discussed in the next section.

3.2.2 Network analysis

This section discusses quantitative analysis of *complex networks* (graphs with non-trivial topological properties), which aims to measure specific properties of the network and its parts that may reflect real-world phenomena. By contrast, visual analysis, which is discussed in the next section, focuses more on qualitative aspects and especially local connectivity. The co-occurrence networks are weighted, which excludes some simple network measures that are defined for binary networks only, or would require a lossy binarization of the network [161]. Connection weighting provides robustness when dealing with potentially noisy input, and it makes especially smaller networks more informative. The methods presented here are selected because they account for the information present in the weighted co-occurrence network, and because they can be applied to undirected networks, while they also may be applicable to text-derived networks other than co-occurrence networks. The quantitative analysis falls into two main categories: *global properties* and *node centrality*.

Global network properties. Measures of global properties provide descriptive statistics of a network and the phenomenon it represents. For instance, real-world networks commonly have a very small average distance among nodes relative to the size of the network, called a "small-world" property [212]. It has a functional justification in making communication over the network efficient relative to the cost of maintaining links and nodes, while it also follows general tendencies in non-regular networks whose nodes have a varying number of links, or *degree*. The degree distribution is central to describing a complex network. Natural networks typically are *scale-free* and follow a power-law, or Zipfian, degree distribution ($P(k) \sim k^{-\gamma}$ for the fraction of nodes $P(k)$ with k links), which reflects their evolution through processes of preferential attachment, i.e., the likelihood of a link in formation attaching to a specific node is proportional to its degree (colloquially also known as "rich gets richer") [12]. Some natural networks also exhibit expo-

nential or hybrid power-law/exponential distributions, reflecting alternate processes of network formation [106, 105]. In order to account for connection weights, degree may be substituted for *strength,* $S(i)$, which is the link weight sum per node [14], and strength distributions calculated.

Beyond single nodes, analysis of network structure may also target the density of neighborhoods and the modularity of a network. Measures such as the clustering coefficient (probability of triplets of nodes being fully connected) quantify structural density at a global level for descriptive analysis, whereas other approaches perform different forms of clustering on the network to identify densely connected regions or communities [72], suitable for exploratory analysis.

Node centrality. In real-world networks where some form of transfer is taking place, node centrality is a property of particular interest, as it reflects the general importance of a node in a network in terms of the influence it may exercise on other nodes, or conversely, how exposed a node may be to the rest of the network. In the present text mining setting, text serves as a data source for deriving networks that approximate real-world structures, e.g., associations among physical entities, and centrality measures thus are assumed to approximate importance of real-world entities. Common measures of node centrality include degree centrality (node degree as a fraction of total number of nodes), and shortest-path-based closeness (average distance between nodes) and betweenness centrality (fraction of paths passing a node). For the latter two, Dijkstra's shortest-path algorithm [63] can be applied also on co-occurrence networks by inversing the weights into distances [156]. Shortest-path-based measures assume flows along optimal paths between nodes, whereas other measures have been proposed, such as *information centrality* [194] (equivalent to current flow closeness centrality [38]), that reflect the free spread of information in networks, not routed, but propagating along possibly parallel paths.

Information centrality with smoothing. The notion of information centrality may be of general interest to a range of text-derived networks: for modeling the systemic importance of banks (see Section 4.1.1), regulatory significance in biological networks [206], thematic centrality of keyterms based on association links (see Section 4.2.2), and potentially many other networks based on chaotic systems. Next, I discuss information centrality in more detail and its application to text-derived networks, and in particular to co-occurrence networks. The information centrality measure is defined as:

$$I(i) = \frac{n}{nC_{ii} + \sum_{j=1}^{n} C_{jj} - 2\sum_{j=1}^{n} C_{ij}} \tag{3.1}$$

$$C = B^{-1},\ B_{ij} = \begin{cases} 1 + S(i), & \text{if } i = j \\ 1 - w_{ij}, & \text{otherwise} \end{cases}$$

for a network of n nodes with a weighted adjacency matrix w and node strengths S. The weights determine how well a signal is carried across the network, such that weaker links result in stronger deterioration.

In experiments, I observe that the measure is quite sensitive to variations in the connected component size of the network. If a weakly connected node disconnects completely from the rest of the network, a single disconnected node will measure zero centrality, while nodes of the main connected component will all measure as significantly less central. This makes data sparsity an issue to comparing centralities between networks (e.g., of different aggregates). Inspired by the use of Laplace smoothing in language modeling [50], and justified by the adherence to the Zipfian distribution shared by language data (term counts) and complex networks (node degrees), I introduce the procedure for the purpose of stabilizing the centrality measure in the face of sparse data. It is an additive smoothing scheme that modifies the weights according to $w'_{ij} = w_{ij} + \alpha$, where $w_{ij} = 0$ for non-adjacent nodes and α is a small smoothing constant (e.g., 1.0). The intuition is that the uniform operation discounts some probability from observed links and saves it for unobserved links, including between pairs of nodes that would otherwise not be connected. This counters the uncertainty of performing limited sampling on sparse data, thereby making the information centrality measure more stable with respect to node pairs being weakly connected or disconnected. The procedure is evaluated in Section 4.1.1 and Paper I.

3.2.3 Network visualization

Accompanying quantitative network analysis, network visualization offers means to explore and understand network structure qualitatively, focusing on the local level of specific nodes, links and neighborhoods, or more global topological features. Networks are often rich in information and, while quantitative measures provide exactness and support in terms of summary statistics, visual analysis provides a more nuanced view of the details. Network visualization draws upon graph drawing as a field that combines graph theory and information visualization. In accordance with the visual analytics perspective, algorithms play a key role in supporting visual analysis of networks, both in visual presentation and in modeling.

In visual presentation, graph drawing algorithms are used to calculate layouts that optimize the readability of the network. Layouts primarily govern the placement of nodes and their types include circular, hierarchical and force-directed layouts. Moreover, the visualization of dynamic networks introduces yet more complexity and room for variation [16]. I discuss here

visualization by force-directed layouts [204, 109], where network dynamics can be visualized in separate frames for chronological aggregates.

Force-directed layouting offers an intuitive view of the structure of natural networks, by placing nodes so that the distances between connected pairs of nodes approximate the inverse of their link weights. This is generally implemented as a physics simulation where links establish attracting forces between node pairs with strength relative to the weight, often with repulsing forces among other nodes as well. Thus, these algorithms group more densely connected nodes together and push sparsely connected parts of the network further apart. The strength of this layout comes from its exploitation of the perceptual design principles of spatialization and connectedness, as discussed in Section 2.3.1, which support intuitive reasoning about groups and their relations in terms of densities and position. Force-directed layouting is a form of multidimensional scaling (cf. dimensionality reduction, Section 2.2.1), where the distances of the network's adjacency matrix that could be perfectly represented in a high-dimensional space are scaled down to two dimensions. The visualized links then counter the distortions by illustrating the similarity relations explicitly, using a less constrained visual mark and channel such as a line (with optional encoding of link strength by intensity or thickness). Still, the layouting has difficulties to maintain readability when scaling to large networks, particularly with high link density and low degree of modularity, e.g., making it difficult to avoid a high degree of line crossings. Because of these constraints, filtering and other types of transformations to the network should generally be considered, in order to support clarity and exploration.

Interactivity is helpful as a way of letting the user manipulate the network layout and other aspects of its representation. Manually moving nodes can ease readability in cluttered areas, while it also allows interaction directly with the layout algorithm so that the user can push the optimization process out of a local optimum in order to explore alternative arrangements. Through this form of interaction, a close-knit collaborative optimization/exploration loop emerges. Interactive network visualization by the D3 force algorithm [36] enables such collaboration and is applied to various types of networks in Sections 4.1-2. Other types of interactive view manipulation can also be applied, for instance, in order to assist exploration of certain parts of the network by highlighting or displaying additional information on demand.

Networks serve as a data representation both for presentation and for further computation and modeling. Algorithms extract and transform information to construct networks, and they refine the information further by performing modeling on the structure. For instance, network visualization may incorporate information such as node centrality, in order to integrate quantitative analysis and thereby support the qualitative, visual exploration.

The basic network structure provides a scaffold where further information, from modeling or from other data sources, can be attached and visualized. Because the network represents a fundamental form of organization, its visual interactive representation functions as a rather generic interface, which enables overview, navigation and exploration of details and related information on demand.

Interpreting language in terms of entities and their relations, the network organization becomes a natural way of representing text for computational processing and for human analysis. This section has touched upon modeling of syntagmatic relations, and in particular expressed relations among real-world entities, as one way to approach text data for understanding their contents. At the same time, the section has discussed networks and their analysis, which is to serve other types to text analysis in what follows.

3.3 Semantic modeling

The relation modeling presented in Section 3.2 focused on syntagmatic and general relationships among entities in a text, without attempting to discern and distinguish the nature of those relationships further. Although it represents a lightweight approach to mapping text contents, requiring very little prior knowledge at that, the relations remain very general and, therefore, often unfortunately difficult to interpret in more meaningful ways. This prompts for addressing meaning, in the contexts where entities are mentioned as well as in text in general, i.e., it prompts for the modeling of semantics. Sticking to the research focus of exploring knowledge-lean methods, this section targets unsupervised learning approaches to semantic modeling, which also are based on notions of term co-occurrence, starting from a topic modeling point of view and moving toward more granular modeling based on word-level semantics, and exploration of associative, paradigmatic relations.

Assuming a distributional semantic approach, representations of meaning are learned for words and also sets and sequences of words. These representations are applied in the next chapter as such, in order to map topics in text (Section 4.2), while they also are used as features for supervised learning tasks (Sections 4.1.2 and 4.3). Semantic modeling is introduced in the following sections, and Section 3.4 places it in the context of predictive modeling.

3.3.1 Probabilistic topic modeling

The knowledge-free nature of topic modeling fits the exploratory assumption that we may know very little to nothing about a body of text to be analyzed, as it seeks to identify general themes in order to provide overview

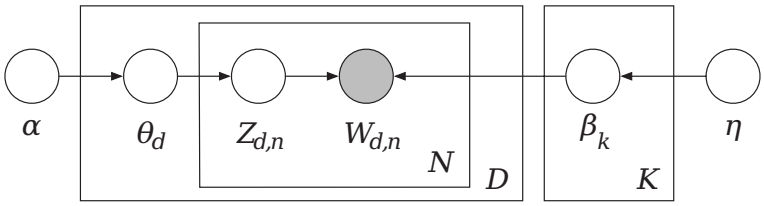


Figure 3.2: Plate diagram of the Latent Dirichlet Allocation topic model.

and organization of a corpus. Topic modeling has had early uses within information retrieval, where methods such as *Latent Semantic Analysis* (LSA) [60] apply dimensionality reduction on a document-term matrix (containing term occurrence counts) in order to obtain a representation of terms in a low number of dimensions (corresponding to topics). *Probabilistic Latent Semantic Analysis* (PLSA) [97] later introduced a probabilistic interpretation of the problem, followed by *Latent Dirichlet Allocation* (LDA) [33] that provided a refined Bayesian approach using a Dirichlet prior distribution, helping to reduce overfitting.

LDA has become widely used, and has inspired numerous extensions such as hierarchical and dynamic topic models [31]. This section focuses on LDA as a practical general-purpose method for studying topics in a corpus. It is a generative topic model that represents the topic structure of a corpus as a set of probability distributions, which provides an interpretable foundation. Nonetheless, the model is rich in information and its interpretability is dependent on meaningful presentation.

An overview of the model is provided in Figure 3.2. Based on the observation of terms $W_{d,n}$ in documents $d \in D$, more precisely their co-occurrence within each document, the model infers a given number of latent topics. LDA rests on the assumption that each document may discuss a mixture of several topics, where a probability distribution θ_d over topics defines the assignments of a document d, and is controlled by a sparsity parameter α (Dirichlet prior). A topic $k \in K$ is described by a distribution β_k over terms, with sparsity parameter η. The model also defines per-term distributions $z_{d,n}$ over topics within a document.

The distributions are useful for interpreting the topic structure inferred from a corpus. The topic-document probabilities of θ_d provide a link from topics to documents that can support information retrieval based on specific topics, while the distribution directly represents an embedding of documents in the latent topic space (of dimensionality $|K|$), and as such LDA constitutes a method for learning continuous low-dimensional representations of sparse and discrete data. Hence, topic modeling may be useful as a tool both for exploration of text and as a dimensionality reduction and representation learning step in a larger machine learning setup.

The topic-term probabilities of β_k reflect the term frequencies within a topic, which can be used to bring forward descriptive keyterms of a topic. For instance, by filtering out function words, the more meaningful words reflecting the content matter remain. Alternatively, terms of different topics may be compared in order to rank terms based on how distinguishing they are of each topic, e.g., inspired by TF*IDF penalizing terms that occur frequently in many topics [32]. In line with Chuang et al. [53], I propose to use the conditional probability $P(k|w)$, as an interpretable measure of how distinguishing the term w is of topic k. The probability of k given the observation of a single term w is derived from β_k (in terms of $P(w|k)$) and $P(w)$ (estimated directly from the corpus) using Bayes' rule [15]:

$$P(k|w) = \frac{P(w|k, \beta_k)P(k)}{P(w)};\ P(k) = \frac{1}{|D|}\sum_{d \in D} P(k|d, \theta_d)$$

In Section 4.2.1, LDA topic modeling is applied and combined with a interactive network visualization for exploration of text corpora, where $P(k|w)$ is used to rank and show the most distinguishing topic keyterms. The use of LDA and the visual representation is described further and evaluated against texts from a patent database.

Separation into distinct topics may itself provide useful organization and classification that supports understanding.[3] However, since topic inference is unsupervised, there is no guarantee for the meaningfulness of topics, and LDA does in fact suffer from some recognized issues of interpretability [46]. While the probabilistic foundation offers interpretable weighting of keyterms and topic assignments, understanding the meaning of and naming a topic based on a ranking of keyterms may be cognitively demanding, especially if the keyterms are semantically incoherent. In practice, topics may be overly broad, narrow or overlapping, which makes it challenging to clearly distinguish and define them. The broadness of topics can be adjusted by the model parameters, primarily the number of topics to infer, and efforts have also been made to incorporate measures of semantic coherence in modeling [147, 155].

Nevertheless, the exploration of topics does not necessarily require such a strict and discrete structure, but instead may also be centered around keyterms primarily, and the notion of higher-level topics more implicitly. Assuming lexical units to be a relatively interpretable atomic unit of meaning, direct modeling of word-level semantics is likely to be meaningful as a more fine-grained approach to modeling semantics and topics. Along these

[3] Compare with the discussion in Section 2.1.1 on the role of classification in language as a foundation for cognition, or Lakoff who states: "There is nothing more basic than categorization to our thought, perception, action, and speech." [119]

lines, the next section discusses distributional semantic modeling as an approach to addressing meaning in text, whereas Section 4.2.2 builds upon it to construct network models that in combination with interactive visualization provide means for exploratory topic modeling.

3.3.2 Distributional semantics and word embeddings

Distributional semantics is an approach to semantic modeling relying on the distributional hypothesis that linguistic items of similar meaning tend to occur in similar contexts [85]. For instance, words whose contexts have similar distributions of words, i.e., they are used in the same kind of contexts, are thereby similar in meaning. The idea is well captured by the aphorism: you shall know a word by the company it keeps [71]. Context is typically more locally defined than in topic modeling, and may focus on word co-occurrences within the span of a few words, a sentence or a paragraph.

Early approaches to distributional semantic modeling have been based on counting of words and comparing their sparse distributions to produce word embeddings [13]. Word co-occurrence can be analyzed within positionally defined contexts, or within structural contexts such as along paths of dependency relations [163]. Different forms of transformation are typically applied to the count distributions, including dimensionality reduction methods such as *Singular Value Decomposition* (cf. [185, 60]). Instead of counting, neural networks have more recently been trained to predict between contexts and target words (e.g., [19]), and in the process learn dense distributed representations of words, akin to the dimensionality-reduction-based approaches. Neural approaches to word vector training have been successful in scaling to large data sets and producing highly accurate results [142, 13], and are practical to incorporate into a unified neural learning framework to serve the purpose of representation learning (see Section 2.2.3).

The applications in Sections 4.2.2 and 4.3 use word embeddings produced by the *skip-gram* model [142] (illustrated in Figure 3.3, and part of the *word2vec* implementation) that learns to predict contexts from a center word using a shallow feed-forward architecture with a linear projection layer, with words presented in a binary format. The model has lower computational complexity than previous recurrent language models, but also than earlier shallow feed-forward language models with non-linearity in the hidden layer. A hierarchical softmax layer and binary Huffman tree coding of words further reduces the computational complexity. As a result, the model can be trained in very practical time on billions of words using regular hardware, while yielding highly accurate vectors. After training, the weights between the input and projection layer are extracted as word vectors (compare with the weight matrix $U^{(1)}$ in Section 2.2.2).

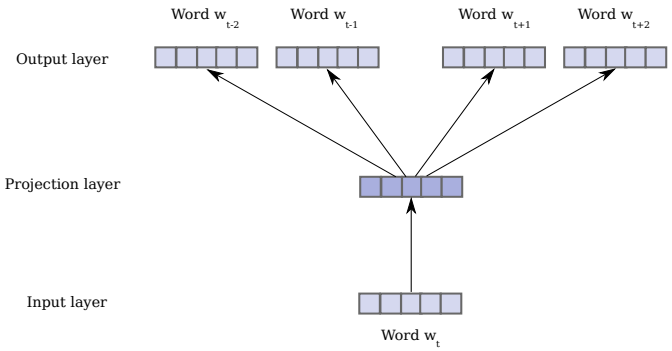


Figure 3.3: The skip-gram neural network model for learning representations of words by predicting context words.

Semantic similarity between words is measured as their distance in the semantic space, e.g., in terms of cosine similarity: $sim(t_1, t_2) = v(t_1) \cdot v(t_2)$ (with unit vectors v for terms t). Mikolov et al. [142, 144] demonstrate that skip-gram word vectors also can support vector arithmetic to exploit regularities in the space that approximate semantic and syntactic regularities well. For instance, they show that the position $v(\text{“}king\text{”}) - v(\text{“}man\text{”}) + v(\text{“}woman\text{”})$ in the space is closest to $v(\text{“}queen\text{”})$. Likewise, this arithmetic operation can also resolve analogies such as Athens is to Greece as Baghdad is to *Iraq* and code is to coding as dance is to *dancing*.

Limitations. Word vectors are highly useful as a data-driven means to compare and organize symbolic input, and, as demonstrated throughout this thesis, they serve an important role as representations for further modeling. They do, however, come with the assumption that words are a natural unit of meaning to operate on. The word level generally may constitute a rather well-balanced compromise in choosing how to dissect a text, as single words may represent a rather atomic unit of meaning that is frequent enough to support aggregation and reliable quantification.[4] Nevertheless, as discussed in Section 2.1.2, meaning is defined both below and above the level of single words, implying that breaking up text into individual words inevitably gives rise to issues of interpretability. The aggregation of occurrences of linguistic units comes at a loss of their individual contexts. We gain a generalized representation of the meaning of that unit, at the cost

[4]It may be relatively practical to aggregate over words in English and other morphologically simpler languages, whereas morphologically rich languages such as Finnish exhibit more variability at the word level and may require additional processing. By contrast, loss of meaning by splitting compound nouns, for instance, may be a greater issue in English than languages such as Finnish, Swedish or German where they are written as one. Additional processing may also be introduced to identify and join multi-word expressions.

of its full interpretability in context. This loss may be detrimental to both subsequent modeling and final interpretability of output. The following section discusses a workaround that, while sticking with words as the input symbols, models words as a sequence rather than as a *bag of words*, i.e., according to the prevalent word-order-agnostic approach assumed, among others, in topic modeling.

Another limitation of word vectors, such as those produced by the skip-gram model, is that their measure of semantic similarity tends to mix both syntagmatic and paradigmatic relations among words. Rapp [172] finds that the two types of relation can be modeled and separated based on co-occurrence in the first and second degree respectively. Moreover, Biemann & Riedl [26] explore graph-based modeling of the two dimensions, which they show to support knowledge-free lexical expansion paradigmatically, guided by the (syntagmatic) context of a sentence. The principle of separating the dimensions could likely be incorporated into dense representations as well, in order to provide practical features for other neural-network-based learning tasks.

Finally, distributional semantic representations may also lack exactness compared to knowledge-based approaches such as WordNet [145], for instance by not distinguishing relations such as synonymy and antonymy. Word vectors are, however, often tuned through end-to-end learning against supervised tasks (e.g., sentiment analysis) to make them better represent aspects relevant for the task.

3.3.3 Sequence embeddings

To the end of tackling the issues of the bag-of-words (or bag-of-vectors) approach raised above, this section extends the distributional semantic modeling of the last section from the level of individual words to sequences of words, allowing for the representation of compositional meaning. This is approached in two different ways. First, as a simple extension to the approach based on feed-forward neural networks, in order to allow it to learn representations of sequences in an unsupervised fashion. Second, by applying recurrent neural networks over word embeddings, so that they can learn internal representations of the sequence in support of supervised task learning.

The first approach was introduced by Le & Mikolov [121] as an extension to the *continuous bag-of-words* model which is similar to, and was proposed together with, the skip-gram model [142] (both are part of word2vec). The *paragraph vector* model (also referred to as *Distributed Memory Model of Paragraph Vectors* (PV-DM) or *doc2vec*), illustrated in Figure 3.4, is as the skip-gram model a shallow neural network and it predicts a target word based on the input of previous word context and a paragraph ID that con-

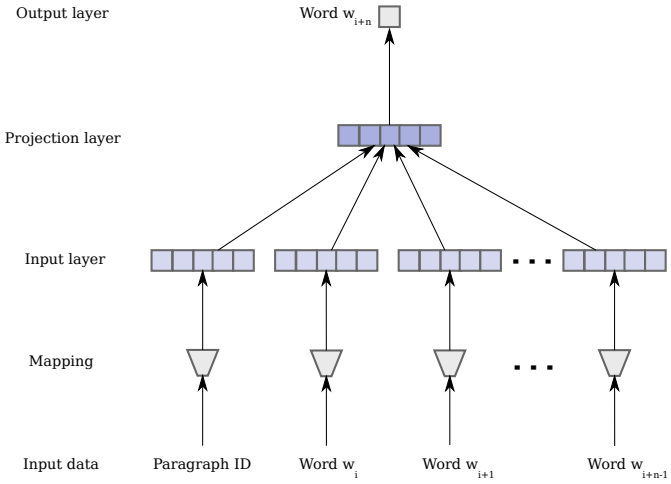


Figure 3.4: The paragraph vector model for learning representations of word sequences.

ditions the task. Paragraph in this context refers to an arbitrary-length sequence of tokens, and as such the model may produce vectors of sentences or documents in the same way. Paragraph IDs are coded similarly to words in the vocabulary and presented as persistent input while the model traverses a paragraph and learns to predict target words. While the word context informs the prediction in the vein of language modeling, adding the paragraph ID allows its weights to capture a global context representation (of the whole paragraph), i.e., capture information necessary for the prediction to differentiate between the particular paragraph and the language in general. As such, the paragraph weight vector functions as a memory that helps inform prediction and thereby capture the semantics of the paragraph. This training procedure results in a single vector that compactly represents the meaning of a sequence of words. This sequence embedding is utilized as a representation for supervised learning as described in Section 3.4.1, and its application is presented in Section 4.1.2. The paragraph vector model is attractive due to its simplicity and speed, whereas some other models have achieved somewhat better results on reference tasks, e.g., by modeling sentences recursively over parse trees [104, 196].

The second approach mentioned is based on recurrent neural network modeling that holds an internal representation while traversing a sequence. This approach is widely used as part of supervised learning tasks, while some unsupervised approaches also have been explored (e.g., [113]). The paragraph vector model is practical as a stand-alone tool for producing sequence embeddings, but suffers from a drawback in that each sequence has to be explicitly represented, which creates a memory overhead that can impede

scaling to very big data sets. The model is also impractical for representing sequences in an online manner. The latter method described in 3.4.2, and its application in Section 4.3.2, incorporates recurrent modeling using Long Short-Term Memory networks (aforementioned in Section 2.2.3) for supervised task learning in the context of discourse parsing.

3.4 Semantic-predictive modeling

The methods discussed up to this point have all been data driven and highly knowledge free, i.e., they assume very little or no prior knowledge about the text material under study, or the analysis task, and as such are well-fitting for open-ended exploration and for becoming acquainted with a new data set or problem. The freedom entailed, however, can also make interpretation more difficult. By gradually constraining the analysis and quantifying certain aspects, the results can become more meaningful in the sense of being more specific, comparable and structured. Starting from a knowledge-free modus operandi, the initial analysis can help successively target and narrow down the process of inquiry. As Tukey expressed in his seminal work on exploratory data analysis: "It is important to understand what you *can do* before you learn to measure how *well* you seem to have done it." [202]

That said, this section steps into the realm of predictive analysis by investigating how setting some constraints can further the text analysis mission, without departing from the exploratory stance altogether. This is done in line with the knowledge-lean philosophy of minimal requirements on encoding and maximum flexibility, for which the introduced semantic modeling serves the important purpose of representation learning for supervised task learning. Predictive modeling is pursued in two different directions. First, I pursue the topic of identifying events in chronological text, based on implicit definitions decoupled from linguistic form, for which it is easy to set up supervision data. This direction of distant (or weak) supervision is applied to the tracking of financial risk and retrieval of descriptive text segments, in Section 4.1.2, based on the method presented in the next section. Second, I explore the application of semantic and predictive modeling toward natural language processing as a means of dissecting and organizing text in a more refined manner, thereby outlining the reach of unsupervised or distantly-supervised text analysis and studying how to pursue directly supervised modeling of language in a knowledge-lean and flexible way. To be precise, the natural language processing focus lies on discourse parsing (Sections 3.4.3 and 4.3), which may provide useful infrastructure for linking sentences and segments of text, e.g., extracted for the description of events, as a foundation for more structured exploration of corpora and for text summarization (as explored by, e.g., [95]).

3.4.1 Event detection and description by distant supervision

This section presents a method for targeting a specific type of real-world event being mentioned in a chronologically-organized corpus, as a way to narrow down analysis while assuming minimal prior knowledge about the nature of the event. An event data set, which specifies pairs of entities and timestamps for when an event has occurred, constitutes the only supervision for this task. As such, the type of event to be modeled is specified implicitly through the selection of this data set, whereas how these events are expressed in text is inferred by the model.

In contrast to the tradition of *event extraction* (cf., e.g., [30, 98]), this approach does not require any linguistic annotation of event occurrences in text nor does it rely on other linguistic resources, and therefore it is fully language independent and also very easy to adapt to new types of event and types of text. Neither does it extract structured representations of events, but instead retrieves segments of text that describe the events, based on distributed representations. In addition to avoiding the need for annotation, this also circumvents the need to define event types strictly, which enables knowledge-lean, exploratory analysis. Compared to the task of *topic detection and tracking* [2, 222], which typically is fully unsupervised, the present method offers specificity toward the chosen type of event, which supports interpretability of the results.

Labeling text by event data. The event data set is linked to the text through the names of entities and the time and date of the event occurrences. Mentions of entities in the text are registered by named entity recognition, as discussed in Section 3.1.1. Named entity recognition is the only step that may involve text-specific feature engineering, but the set of entities is limited by the event data set and variation in entity names and aliases is itself limited, which poses little issue regarding language portability. The recognized entities are used to cross-reference the contexts where they occur (text segment $s \in S$) using timestamps d_s (inherited from the containing document) against the timestamps d_e of the event data set, in order to cast entity occurrences $e_{s,b}$ as presumed event *related* (1) or *unrelated* (0) with an event involving entity b, or alternatively as *ambiguous* (undefined). The label is defined by an inner (W_{in}) and outer window (W_{out}) in time (where $W_{in} \subset W_{out}$ holds for the intervals), according to:

$$e_{s,b} = \begin{cases} 1, & \text{if } d_s - d_e \in W_{in} \\ 0, & \text{if } d_s - d_e \notin W_{out} \end{cases}$$

The time windows may be specified either to support forecasting of events by targeting text that may contain anticipatory signals, or to support nowcasting by targeting coinciding discourse. The method is evaluated

by the application in Section 4.1.2, which focuses on the latter case. News reporting likely contains most relevant information around the time of the event, making it easier to predict coinciding events and to retrieve informative descriptions based on this data, whereas other types of text, such as forward-looking reports and opinionated material, may be better suited for forecasting (see, e.g., [159]). Thus, in the case of news text, the procedure casts each entity occurrence and its text segment as likely to discuss the modeled event, not likely, or ambiguous (i.e., as *coinciding* (1), *non-coinciding* (0) or undefined), based on the assumption that a recent event is likely do be prevalent in the discussion mentioning an entity that was involved. This produces labels $e_{s,b}$ for each pair of segment s and entity b occurring in it, providing the data $(s, b, e_{s,b})$ needed for modeling.

Predicting events with sequence embeddings. The paragraph vector model introduced in Section 3.3.3 is used in this context to learn representations of token sequences such as sentences or documents, in order to serve as features for predictive modeling of events. Feeding a sentence as input, the semantic-predictive model will provide a probability of the event based on the statement contained in that sentence.

The model, depicted in Figure 3.5, is a feed-forward neural network consisting of an input layer, an embedding layer, a hidden layer and an output layer. The input layer encodes sequence IDs s using a one-hot representation, while the embedding layer holds the corresponding pre-trained vector representations. These are fed through the hidden layer with a non-linear activation function to the two-node softmax output layer, in order to predict the occurrence of an event, as labeled by $e \in \{0, 1\}$. For sequences S and corresponding event label, the learning objective is to maximize the average log probability:

$$\frac{1}{|S|} \sum_{s \in S} \log p(e_s|s)$$

The network is trained by backpropagating errors [179] from the output nodes to the weights connecting to the hidden and embedding layer. Training of the weights between input and embedding layer would not help to generalize across sequences, due to their unique IDs, therefore, these weights are kept fixed during supervised training. Further implementation details of the predictive model are discussed in Section 4.1.2.

Aggregated event index. As each mention of an entity provides some clues toward whether an event is occurring, aggregating these signals can provide a more robust way of detecting events. To this end, event indices are defined to aggregate the posterior probabilities of the model for each

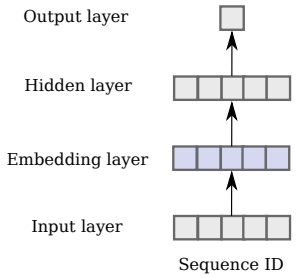


Figure 3.5: Model for predictive modeling of events from text segments. Training of the embedding layer is unsupervised.

instance of an entity occurrence during a given period of time. While the posterior probability $M(s) = p(e_s = 1|s)$ of the neural network model is a measure of relevance to the event for a particular sentence, the entity-level aggregated index measures generally how likely the entity is to be involved in an event at a particular time. The entity-level index $I : p \times b \rightarrow [0, 1]$ is defined as the mean posterior probability:

$$I(p, b) = \frac{1}{|S_{p,b}|} \sum_{s \in S_{p,b}} M(s)$$

for sequences $S_{p,b}$ that contain a mention of entity b and pertain to period p (defined, e.g., monthly). The index serves to provide an overview of event-related discourse over entities and time, while it also provides a basis for evaluation of the model's predictive performance at a more intuitive level of aggregation, namely that of the original event data. Evaluation based on the entity-level index is discussed in conjunction with the application of the model to bank distress events in Section 4.1.2. The signal is aggregated at various levels relevant to the application.

Describing events. A main advantage of text over other types of data for predicting events is the rich descriptive detail it provides. The output of the predictive model is not only a signal to reflect the probability of an event, but it also reflects the informativeness and relevance of the input text to the event. Thus, by using the posterior probability as a relevance measure to retrieve text segments (e.g., sentences) as excerpts, the model can provide informative text descriptions that may offer deeper insight into the underlying developments relating to the event. The excerpts also provide transparency and means to validate the predictive model due to the interpretable nature of the input.

Excerpts can be explored as such, organized in terms of entity and time, allowing a user in the spirit of visual analytics to navigate the descriptions

by the quantitative model output. Interesting patterns of the index time series guide the user toward specific periods and entities, and top-ranking excerpts can then be selected. In order to provide a bit more context to an excerpt and support its interpretation, I propose a derived score for ranking, which can combine posterior probabilities of adjacent segments. The score for an expanded excerpt centered around sequence S_i is formulated as:

$$x_i = max \begin{cases} M(S_i) \\ M'(S_{i-1}, n) \\ M'(S_{i+1}, n) \end{cases}$$

where M' represents an alternative mode of operation where new embeddings are inferred for previously unseen sequences by the paragraph vector model.[5] As the inference is stochastic, the mean vector over n samples is used to estimate the embedding. The rest of the model operates normally to provide the posterior probability. The excerpt expansion may also extend further, to multiple steps away from S_i, if more context is deemed necessary. Visualization of the index and associated descriptions is demonstrated in Section 4.1.2.

Limitations. Inherent to the visual analytics framework is a choice of how to balance the workload between user and computational processing, i.e., the less modeling is performed computationally the more is expected of the user. Greater freedom to read and understand the segments in context comes with higher cognitive demands to gain an overview of the whole breadth of the material, and as a result a user may not be able to effectively take into account all relevant excerpts even with a well-designed visual interactive interface, but rather may only be able to scratch the surface of the material.

The question arises whether further modeling of the descriptions could help alleviate the burden and allow the user to form their understanding based on a broader coverage of descriptive text segments, i.e., if the text could be better organized and summarized. The predictive model already performs one step of extractive summarization by ranking text segments, and additional unsupervised measures may be considered to reduce redundancy among segments in a summary (e.g., Maximum Marginal Relevance [42]) and to find thematically central segments (e.g., TextRank/LexRank [140, 68]). Alternatively, a topic modeling approach as discussed further in Section 4.2 may be employed in order to gain an overview and starting point for exploration. Both approaches can be pursued in an unsupervised,

[5] Le & Mikolov [121] refer to this as a secondary "inference stage", and the functionality has been implemented in *gensim*: https://radimrehurek.com/gensim/models/doc2vec.html

knowledge-free fashion, but both risk causing some loss of context and reduced interpretability. Breaking up text into terms as in topic modeling results in a severe loss of context, where linking back to original contexts to some extent may counter the problem. What the most suitable unit of analysis may be, is a question that bares revisiting throughout the process of developing a text analysis tool. Extractive summarization retains context within the segments, but, without accounting for text cohesion explicitly, it still risks loosing important information from its extended context.

At this point we may be approaching the limits of how text analysis can be effectively and meaningfully supported, without explicit modeling of language. Encountering this limit and recognizing a need for natural language processing in order to transgress it, the question, in the context of this thesis, becomes how this too might be pursued in a knowledge-lean spirit, i.e., requiring as few linguistic resources and maintaining as much flexibility as possible. Although many natural language processing tools already exist, and in many practical cases may be able to answer the need raised above, the discussion that follows considers the development of new tools and how this can be done in accordance with the representation learning paradigm.

3.4.2 Resource-lean discourse parsing

Natural language processing has traditionally relied heavily on linguistic resources for modeling of language phenomena. On the one hand, resources take the form of annotated text, which encodes linguistic knowledge implicitly and serves as data for supervised learning. On the other hand, dictionaries, ontologies and similar resources encode knowledge explicitly and provide features that capture rich linguistic information. While annotation of text to mark the occurrences and structure of linguistic phenomena appears vital, representation learning (as introduced in Section 2.2.3) proclaims that other resources to support feature engineering may not be necessary to linguistic modeling. On the contrary, learned representations are meant to be more flexible, less labor intensive and provide better coverage as they are data driven. Striving for minimal use of manually-crafted linguistic resources may be referred to as a resource-lean approach to natural language processing.[6]

This section presents two neural network models that provide an idea of how resource-lean, representation-learning-based natural language processing may be implemented, as they focus on parsing of discourse structure. Discourse parsing may aid in the above introduced task of describing events. The extracts representing rather large units of meaning, such as sentences,

[6]Knowledge-lean natural language processing would be a suitable term, but I choose to use resource lean in this context, where annotated data is used for supervision and the focus is on avoiding rich linguistic resources for features.

can become more informative through additional modeling of their cohesive relations in the text, i.e., by analyzing how they form part of the discourse of a news article. Discourse parsing, thus, is a form of natural language processing that may support text mining by relating and organizing units of meaning above the syntactic level and across sentences.

The modeling focuses on the more particular case of *shallow discourse parsing*, which follows a formalism that defines flat discourse structures, where segments are connected by a typed discourse relation [171]. The relation type (the *sense*) between two segments (referred to as *arguments*) stand in a discourse relation to each other. Some relations have *explicit* connectives (e.g., *however*, *for example*, *because*), while *implicit* relations lack them. For example, consider the following two adjacent statements:

> "But the market turmoil could be partially beneficial for some small businesses"
>
> "In a sagging market, the Federal Reserve System might flood the market with funds, and that should bring interest rates down"[7]

The relation sense between these discourse arguments is labeled as *Contingency.Cause.Reason* and, lacking a connective such as *since*, it is implicit. By contrast, the initial *but* is a connective that indicates an explicit relation between that segment and a preceding one.

The task of discourse parsing has traditionally received less attention than modeling at lower linguistic levels, and the problem is challenging, as discussed in Section 2.1.3. In particular, the implicit discourse relations are difficult to parse. The following discussion provides a brief introduction to the topic, while Section 4.3 presents the application of the models for multilingual discourse parsing in the context of the CoNLL 2016 Shared Task [221].

Until recently, successful machine learning methods for parsing implicit shallow discourse relations have made extensive use of rich linguistic features (e.g., [170, 102]). While such efforts explore the use of a range of linguistically motivated features, no particular type of feature seems to stand out from the rest, nor is it clear that this laborious approach to modeling is the most effective. In 2015, a few efforts first explored the use of learned dense representations for implicit sense classification for shallow discourse relations [39, 160, 225], and in 2016 several continued on this path (e.g., [214, 180, 49, 129, 107], and Paper V). The methods presented below similarly explore whether representation learning can replace the use of rich linguistic features altogether. Two types of neural networks are presented

[7] The passages stemming from the Wall Street Journal are found in the Penn Discourse Tree Bank training set as relation #31999 (see Section 4.3).

in the following for classifying senses of implicit shallow discourse relations. The models function as key components in a discourse parser, together with other components for argument detection and recognition of explicit relations based on present connectives.

Feed-forward network on bags of vectors. The first model uses a feed-forward neural network topology, which is easier and faster to train than, for instance, recurrent networks. The number of hyperparameters and training time have a notable effect on how throughly the model can be optimized, which means that a simpler model that is better optimized may be competitive with more complex models. In a feed-forward network, the inputs need to be presented in a consistent manner, which in the case of variable length input (variable number of tokens, or other features), or input of variable internal structure, requires transformation to a more invariant form. Therefore, aggregation functions are applied to produce suitable representations of the input sequences. Word order within arguments is disregarded and as such the model assumes a bag-of-words approach, which in this case, operating on word embeddings of tokens, may also be called a bag-of-vectors approach. The embeddings are pre-trained by the skip-gram method introduced in Section 3.2.2.

Traditionally, word pairs retrieved from the respective arguments has been a popular feature, and Braud & Denis [39] show that their dense representation outperforms raw pairs of tokens. The notion of word interactions being distinctive for the task guides the choice to represent arguments separately in this model, as it should support the network in modeling the interactions, based on the latent dimensions in the distributed representation. In a high-dimensional vector representation important aspects of the words are here assumed to be retained well enough, even after aggregation. Following Mitchell & Lapata [151], the aggregation combines the average and pointwise product of vectors. While vector average is generally considered a standard way of combining embeddings, multiplication allows the assumed independent latent dimensions to scale according the mutual relevance among words, which may accentuate some useful information over the coarser averaging operation. The aggregated argument vector is defined as:

$$\mathbf{v}'(j) = \frac{1}{k(j)} \sum_{i=1}^{k(j)} V(j)_i + \prod_{i=1}^{k(j)} V(j)_i$$

for arguments $j \in \{1, 2\}$, where $\prod$ applies $\odot$ (pointwise product) over vectors in $V(j)$ for $k(j) = |t(j)|$ number of tokens.

Figure 3.6 illustrates the described procedures for producing argument representations from tokens. These argument vectors then serve as input to a feed-forward network with a single hidden layer and a softmax output

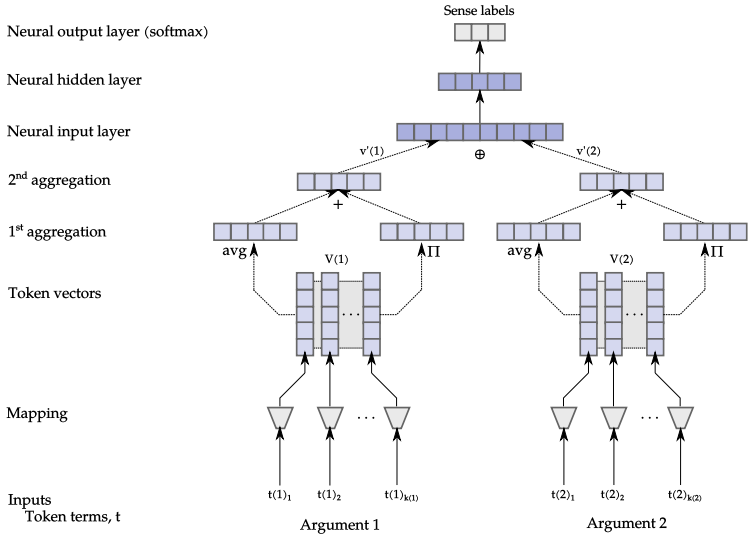


Figure 3.6: Feed-forward neural network and argument representation construction process for the task of implicit discourse parsing.

layer for classifying among sense classes for the discourse relation. The implementation details are further discussed in Section 4.3.1.

Recurrent network on sequences of vectors. Continuing the focus on the above task, the model described in the following takes a different stance by considering tokens as a sequence in order to classify the sense. Word-order-aware recurrent modeling is also linguistically motivated, as it seeks to account for structural aspects of semantic compositionality and the sequentiality of language. It may better reflect the online fashion language users operate in, by continuously relating structural units to previous ones, which van Dijk [205] describes as fundamental to understanding discourse. The feed-forward architecture is relatively simple and practical to train and optimize, while disregarding word-order information may restrain learning.

As an alternative approach, an *Attention-based Bidirectional Long Short-Term Memory* (Att-BLSTM) network is introduced. It is a neural network that extends the representation learning capabilities compared to the previous, and it is applied over a joint sequence of discourse arguments. Argument spans and potentially other information (such as context tokens and connectives) are marked in the sequence by tags (<ARG1>, </ARG1>, etc.), which provides a flexible foundation for modeling. This way of modeling relations in token sequences is inspired by Zhou et al. [227], who use it to model entity relations. Recurrent modeling of shallow discourse relations has been explored by [214, 49, 128], although with separate representation of the two arguments.

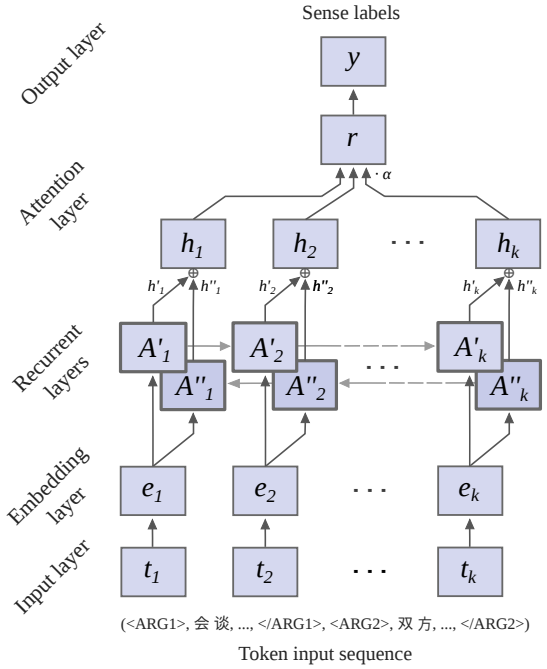


Figure 3.7: Attention-based Bidirectional Long Short-Term Memory (Att-BLSTM) network for classification of shallow discourse relation senses.

The model, illustrated in Figure 3.7, consists of an input layer encoding tokens by a one-hot representation, an embedding layer for their distributed representation, two recurrent layers for sequential modeling of tokens in different directions, an attention layer for aggregation over the sequence, and an output softmax layer for classification among sense labels. Adjacent layers are fully connected to each other in a feed-forward fashion. The network is end-to-end learned, meaning that the token embeddings are updated through backpropagation during training against the sense labels, while they also benefit substantially from being initialized by pre-trained word embeddings. The recurrent layers employ Long Short-Term Memory cells (A' and A'') in a bidirectional mode of operation (called the Bidirectional LSTM [77]), traversing the sequence both in the forward and backward direction, so that both left and right-hand-side contexts can be considered at each step. The LSTM holds a hidden state at each step, represented by a vector it serves as part of the input to the next recurrence step and as output to the next layer. The bidirectional hidden state vectors h_i are combined from the unidirectional LSTM outputs $h_i^{'}$ and $h_i^{''}$ by either concatenation or sum (the operation is here universally denoted by $\oplus$). Before classification at the output layer, the hidden state vectors over the sequence need to be aggregated into a single vector. The simplest solution is to use only the last

hidden state, which in the bidirectional case equals to $h_k^{'} \oplus h_1^{''}$. This vector can potentially hold all relevant information from the traversed sequence for the classification task, but other aggregation techniques may often provide better support for the classification, such as max or mean pooling of vectors, or by the use of attention mechanisms [10].

Attention mechanism. Attention, in this context, is a mechanism for dynamically assigning importance to certain parts of the input, by means of a weighted average of the hidden state vectors, as formalized by Hermann et al. [89]. The attention layer calculates a weight vector α (of length k) based on a trainable vector w (equal length to h) and the hidden state vectors H (consisting of $h_i = h_i^{'} \oplus h_i^{''}$), as:

$$\alpha = softmax(w^T tanh(H))$$

$tanh$ denotes the hyperbolic tangent function, a common activation function. The attention layer uses the dynamic weighting α to produce the aggregated vector:

$$r = H\alpha^T$$

The benefit of using attention, apart from potentially improved predictive performance, is that the weighting provides means for interpreting how the model arrives at its classification decision. Neural networks, and deep networks in particular, are known to be opaque *black box* models. The lack of transparency may discourage their use in many data analysis settings despite their strong predictive performance. As mentioned above, the task of implicit shallow discourse parsing suffers from rather limited understanding of what the distinctive features in a pair of arguments are. Thus, an interpretable model for discourse relation recognition stands to benefit the linguistic understanding of the problem in general, as well as text mining tasks that focus on the text in particular.

In Section 4.3.2, this attention-based recurrent model is applied to recognize discourse relations in Chinese and the α weighting is used to visualize the influence of parts of the input sequence. This demonstrates how machine learning can serve as an integral part of the visual analytics framework, not only through the visualization of model output in the classical sense, but it shows that visualization can serve also a purpose of opening up and making the inner workings of advanced models tangible.

Chapter 4

Applications

"All models are wrong but some are useful"

– G. E. P. Box (1919-2013) [37]

In this chapter, the introduced methods are applied to solve concrete problems. The first section explores their application in the domain of systemic financial risk, in order to model discussion related to banks. Section 4.2 assumes a more general approach to model and visualize themes in text corpora in a fully knowledge-free manner, which is evaluated on patent texts. Finally, the chapter discusses the application of the introduced models for recognition of discourse relations to English and Chinese text.

While the applications share many methodological aspects, the disposition here has a clear mapping to the papers: Section 4.1 reflects the work of Papers I-II, Section 4.2 of Papers III-IV and Section 4.3 of Papers V-VI. The discussion below reviews the general problem, data and results of each of the applications.

4.1 Systemic risk analytics on text

Problem. Computational methods for measuring systemic financial risk has gained substantial interest following the global financial crisis (cf., e.g., [28]), which started to unfold in 2007-2008 in the banking sector and has had persistent effects on the real economy, society and democracy [73, 118, 116, 168]. The massive negative impact has spurred interest to develop means to better understand the financial system, in order to prevent similar destructive events from reoccurring. Meanwhile, the dissemination of more advanced machine learning practice into the field of economics and the potential of big data provide fertile ground for the development of new tools to this end.

At heart of the crisis were major events of bank distress that had *systemic* effects, i.e., failures created shocks that spread internationally throughout the banking system, and affected the global financial system as a whole [118]. The analytics approach to safeguarding financial stability has involved a particular focus on constructing network models that quantify interdependencies among banks in order to understand the system. These networks are based on numerical data on lending and payment flows between banks, as well as indirect linkages based on co-movements in market data [44]. Build-up of risks in banks has also been modeled based on accounting data (see, e.g, [55, 134]).

Systemic risk analytics suffers from a general data problem, in that access to data that directly captures imbalances and interconnections is highly restricted, in many cases even for regulatory bodies (cf. Paper II). Moreover, low reporting frequencies and long publication lags are commonplace, for instance, for accounting and macroeconomic data. Market data offers a timely and widely available source of information on stress, volatility and co-movements [83, 146, 48], but does not carry descriptive information. In this context, text data constitutes a complement that is semantically rich and able to directly describe signals and structure extracted through its modeling, while also being timely and abundantly available. In the spirit of data mining (see Section 1.1.2), and in times of big data and data science [61], there is new interest within systemic risk analytics for taking new types of data, repurposing and utilizing them in novel ways [100].

The present focus on utilizing the unstructured information in text for modeling and understanding bank interconnections and distress represents pioneering work in this direction. Some previous work has utilized text for the identification of risks, e.g., in the financial domain (cf., [41, 35, 192, 209, 99]) and for crisis monitoring regarding violent and disaster events (cf. [197]). Nyman et al. [159] show that shifts in analysts' sentiment were detectable ahead of the financial crisis, as a noteworthy example of sentiment analysis targeting systemic risk.

Data. In this work, both the construction of bank network models and modeling of distress events rely primarily on news text. The data is obtained from Reuters online archives and spans the period from 2007Q1 to 2014Q3. The data set contains 6.6M articles in total.[1] An early version of the work on bank networks,[2] applied similar modeling on discussion in Finnish from an online discussion forum, which demonstrates that it is readily applicable to new languages and types of text.

[1]The experiments in Paper I use a 45% random subsample available at the time.

[2]See list of other co-authored publications, paper nr. 4.

Specifying the recognition of named entities is the only configuration step required, which may involve both specifying entities and language patterns for all naming variants. In the reported experiments, named entity recognition was performed by manually developed regular expression patterns that cover the set of banks, with common aliases, spelling variations, abbreviations, etc. There are 27 European banks under study in Paper I, many of which are considered systemically important, and in Paper II a longer tail of banks that have suffered distress was included to reach a total of 101.

Results from network analysis are also compared against accounting data of the banks (see evaluation in Paper I). The modeling of bank distress is based on an event data set consisting of bank names and dates when they have been declared distressed (243 instances). The data originates from the European Commission[3] through the European Central Bank [22], and defines the events to include government interventions, state aid, direct failures and distressed mergers.

4.1.1 Bank networks from text

Paper I studies bank interconnectedness based on news reporting. The method for constructing co-occurrence networks, described in Section 3.2.1, is used to produce quarterly aggregated networks among 27 major European banks. Methods for quantitative network analysis (Section 3.2.2) and interactive visualization (Section 3.2.3) are used to study the characteristics of the network, and its dynamics over time.

Results. Figure 4.1 shows a snapshot of the interactive visualization at 2008Q4, illustrating the connectivity in reporting among banks in the months immediately after the market crash in September 2008. The period saw a general increase in reporting about banks and a strengthening of the links in the co-occurrence network. The network is densely connected, with a very strongly connected core of major banks, the majority of which are classified as Globally Systemically Important Banks (orange nodes). The notion of node centrality is quantified in terms of information centrality, visualized both as node size and in more detail in Figure 4.2. Centralities peak especially in late 2008 and early 2009, following the crash, as well as in early 2012, which coincides with concerns over the spreading Eurozone debt crisis.

The information centrality of banks corresponds well to their positions in the network visualization, as optimized by the force-directed layout algorithm, and generally reflects the systemic importance of the banks. Further evaluation presented in the paper shows that the centrality measure is highly correlated with bank size. It does not measure vulnerability directly, but

[3] See http://ec.europa.eu/competition/state_aid/overview/index_en.html

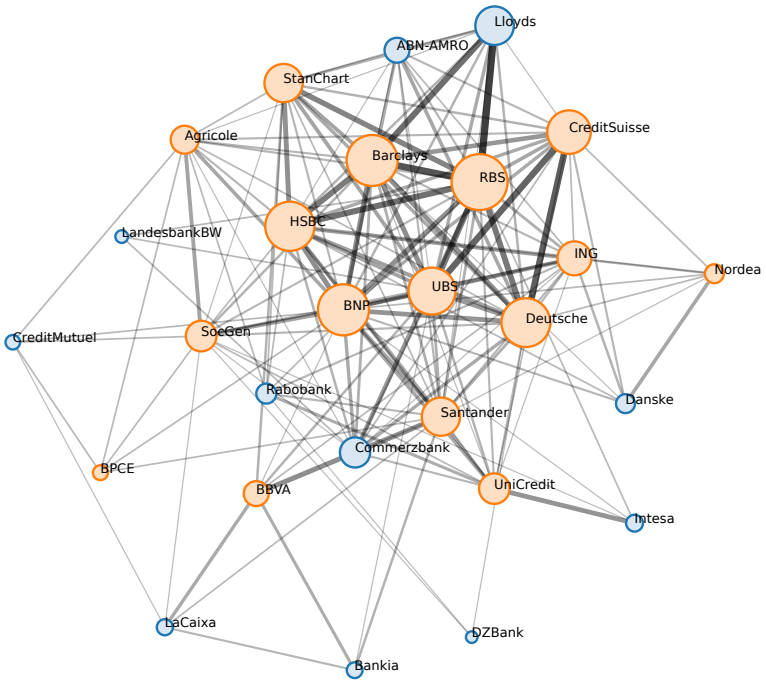


Figure 4.1: Visualization of the co-occurrence network for major European banks in the aftermath of the financial crash (2008Q4). Node size indicates information centrality of banks, and link opaqueness and width co-occurrence frequency. Nodes classified as Globally Systemically Important Banks in the data are colored orange.

rather provides a broader measure of interconnectedness that, due to the descriptive detail of text, can be further narrowed down through semantic modeling.

The time series in Figure 4.2 show the centralities of banks over time with different levels of smoothing. With light smoothing, interesting global patterns are accentuated, such as the peak starting in late 2008. Without smoothing, changes in connected component size has a strong influence on the centrality measurement of core nodes. Stronger smoothing reduces global fluctuations and highlights dynamics in relative node centrality, thus supporting comparison among banks.

This work represents one of a few pioneering effort of text mining within systemic risk analytics, and, to my knowledge, it is the first to utilize text for this problem. It has been cited in the Bank of England handbook *Text mining for central banks* [23], where they state that text mining has been infrequently used in economics and particularly within central banks, while it would be a "useful addition to central bank's analytical arsenal".

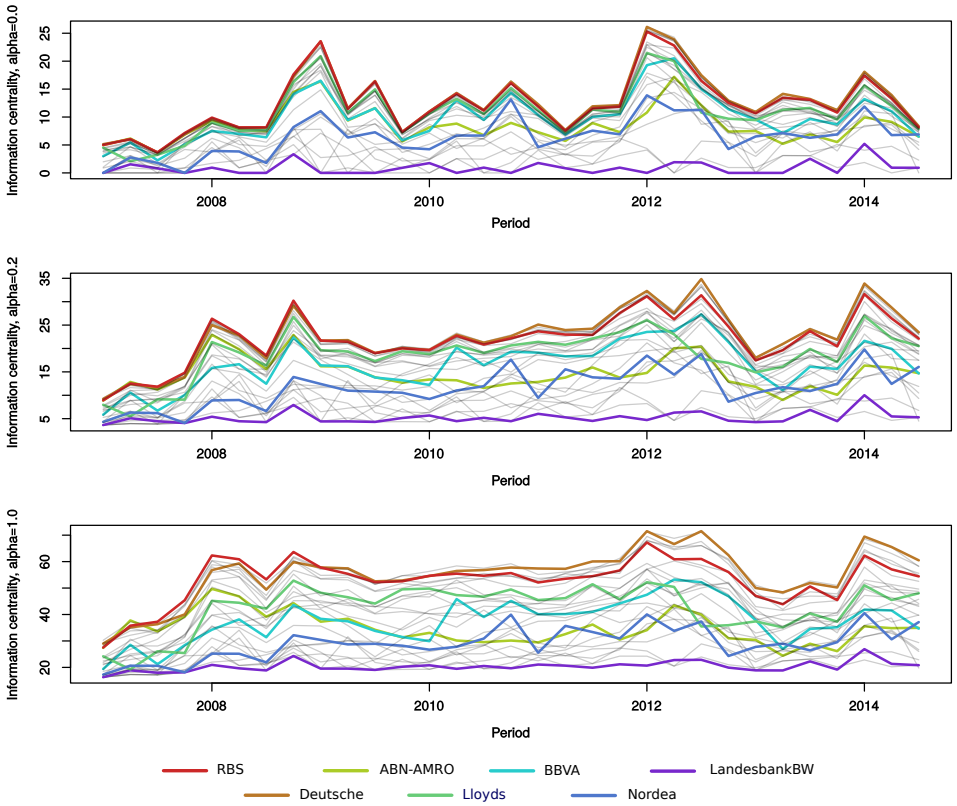


Figure 4.2: Information centrality of European banks in news reporting over time, with different levels of smoothing. Selected banks are highlighted.

4.1.2 Detecting and describing bank distress by news

The co-occurrence network approach is simple, flexible and knowledge lean, as only the recognition of bank names requires background knowledge. At the same time, the generality limits its interpretation, and narrowing down the analysis is motivated from a domain perspective, in order to produce more concrete and meaningful results. The work presented in Paper II moves from the general overview of the bank landscape to a direct focus on bank distress, by applying the method for modeling of events described in Section 3.4.1.

Results. The method takes a knowledge-lean approach to event modeling that is suitable for exploring this niche application. The news articles are processed sentence by sentence, namely by scanning for bank mentions and by learning representations for each sentence. 716k sentences are registered as mentioning any of the target banks, and after cross-referencing with the event data set 386k sentences are labeled as coinciding/non-coinciding with

a distress event of the bank. The rest are either ambiguous or occur outside of the event data set that spans the period 2007Q3-2012Q2, and are not used for training or evaluation, only in deployment.

The neural network model is optimized to an embedding size of 600 and a hidden layer size of 50. An optimized threshold on the index $I(p, b)$ is used to classify between distress and tranquil states for bank b in period p (aggregating over posterior probability for an event $p(e = 1|s)$). The model is evaluated based on a skewed preference for false positives over false negatives, to reflect the intended use of the model in a supervisory setting where missing a crisis event is very costly, whereas a sensitive model is able to alert on potential dangers and initiate further investigation. The evaluation metric, called relative Usefulness U_r, incorporates this preference (set at 9:1). Unlike the in text mining popular F-score, it also measures the gain over majority class (tranquil) prediction that due to a highly skewed class distribution would achieve 91% accuracy on the data. The task is challenging as a model has to surpass the 91% accuracy threshold in order to perform better than an uninformed decision, i.e., in order to measure positive Usefulness.

The evaluation, described in detail in Paper II, shows that the model is able to effectively classify bank distress events from news, reaching U_r of 32.6% with random sampling and 12.3% using a more conservative leave-N-banks-out sampling scheme (a perfect model has $U_r = 100\%$). The mean posterior probability I that aggregates over multiple sentences per month and bank demonstrates more stable results compared to the raw sentence-level predictions. Although the experiment is not strictly comparable to related work due to differences in settings, for instance, Betz et al. [22] survey methods on conventional data that achieve U_r of 19-42%, which may serve as a reference. To conclude, the news-based model achieves decent predictive performance but does not surpass conventional models. This indicates that there is value in using text-based information sources for the traditional detection task, while it is likely that it best serves as a complementary input together with numerical indicators. The evaluation validates the quality of the predictive model, which, in the present context, has the important function of providing an assurance about the quality of the text descriptions the model provides.

Applied to the whole range of news articles, the model outputs 716k posterior probabilities distributed over 93 months. The aggregated index I reduces the data, and Figure 4.3 visualizes the distribution in the monthly cross sections. The blue line represents the mean distress probability in Europe and the gray lines show every 2.5^{th} percentile to give a clearer impression of how the stress levels evolve. The percentiles provide a more structured view and highlight interesting patterns, e.g., following the crash in September 2008 a majority of the banks in the cross sections peak in

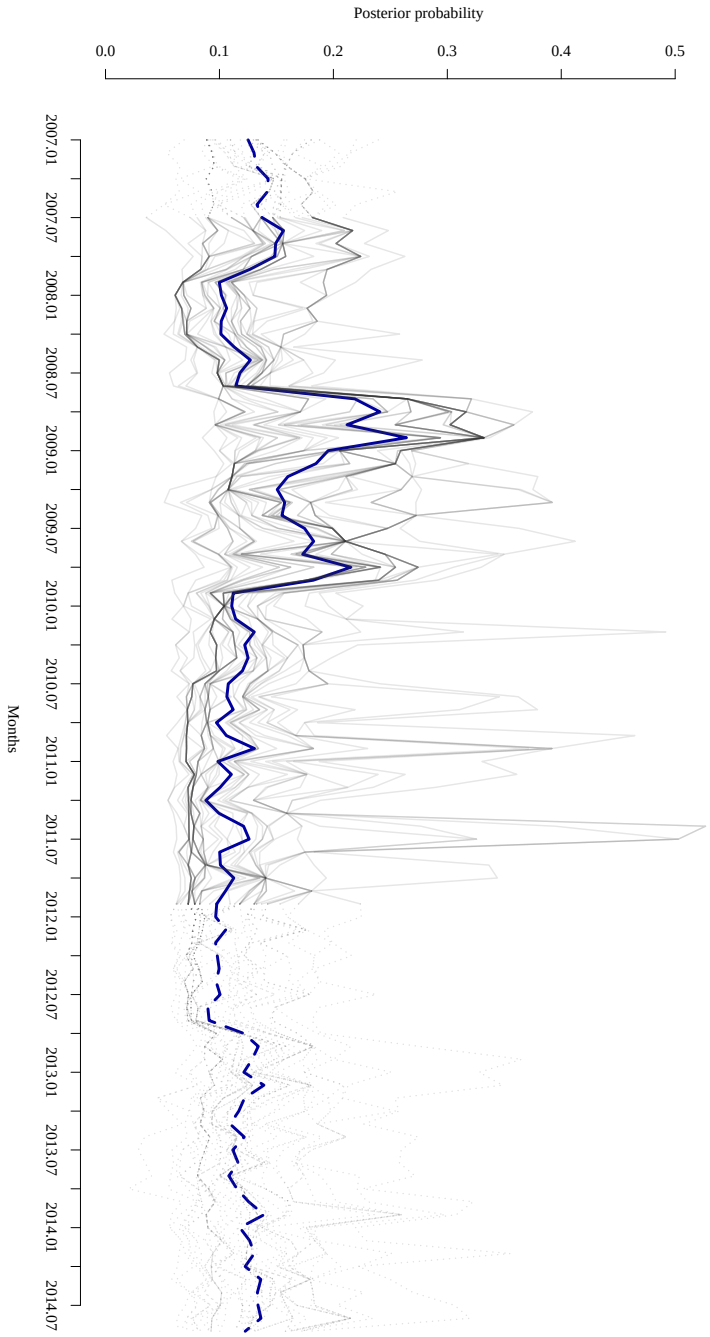


Figure 4.3: Levels of distress-related reporting in the news over time. The mean probability over all banks is shown in blue and the distribution is plotted as percentiles (at every 2.5%) in gray. Dotted lines show model output when deployed beyond the time span of the original event data set.

distress-related reporting, and later, as the mean stress decreases, strong eruptions among a few banks continue to occur. This reflects a general concern about bank distress in winter 2008/2009, whereas single countries and banks experienced concrete problems for a long time to come.

This form of overview provides means for exploring and focusing on specific parts of the data, in order to better understand the developments and the phenomenon. Figure 4.4 shows a similar type of view, visualizing country-specific mean stress levels for pinpointing interesting patterns and retrieving descriptive excerpts from the underlying news reporting. The figure includes the examples of Belgium and Ireland, with key points marked and interesting excerpts included. The excerpts are selected from the 10 highest-ranked excerpts for each period and country. Qualitative analysis, discussed in more depth in Paper II, shows that the model is able to find highly relevant excerpts that can be browsed to study the developments of European bank distress.

The present work presents the excerpts as independent descriptions, consisting of a sentence mentioning the bank and a context sentence on each side to support interpretation. The model could support an interactive system for exploration similar to Figure 4.4, where the user is allowed a lot of freedom to explore points on the curves and associated rankings of excerpts. Such an approach, however, would place much responsibility on the user to relate the excerpts and deal with redundancies. Thus, it is motivated to not only focus on the interactive interface going forward, but also on language processing methods for structuring and summarizing the excerpts. The next section takes a closer look at visual interactive design for exploration of semantic similarity structures in text, which could be recombined for this application. Likewise, the discourse parsing discussed at the end of the chapter is a concrete form of natural language processing that could introduce structure among extracted sentences, based on their place in the discourse of an article, rather than only based on semantic similarity among sentences.

The predictive model could also readily be used in co-occurrence analysis to produce more strictly defined relations that reflect when banks are mentioned together in a distress context. Networks built from such relations would be easier to interpret, and would possibly better reflect systemic vulnerability of banks. Parallel to the more exploratory approach to studying text descriptions, this would represent a direction closer to traditional forms of systemic risk analytics based on conventional types of data.

This work has been featured as an example on how text analysis can serve as a novel approach in the supervision of financial risks by the Swedish Riksbank in their commentary on future information supplies for central banks in the light of big data [100]. The model's ability to describe the events

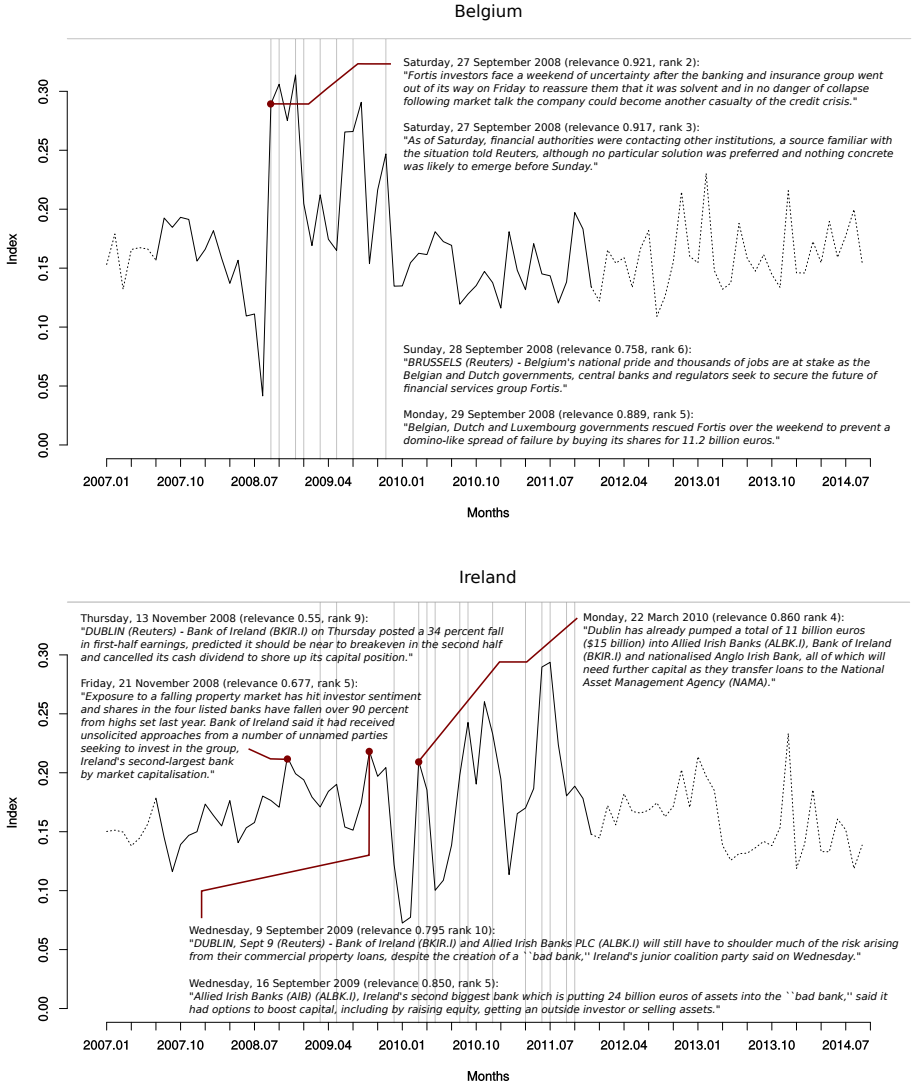


Figure 4.4: Mean distress over time in Belgian and Irish banks with extracted descriptions at key points of the curve. Vertical lines indicate when recorded distress event occurred.

that are being detected, as highlighted by Bloomberg,[4] has also received more wide-spread attention.[5] Moreover, interested in the knowledge-lean approach to modeling of events, the work has been replicated by Thomson Reuters.

4.2 Visual topic exploration

Problem. This section explores an application that is not tied to a particular domain, but rather offers means for fully knowledge-free analysis of the contents of a body of text. This assumes a topic modeling approach to dissecting the thematic composition of the material, using both Latent Dirichlet Allocation topic models (LDA; Section 3.3.1) and word-embedding-based modeling (Section 3.3.2). While methods for topic modeling, including LDA, are readily available, they produce rich information but are not necessarily easy to interpret as such (see discussion in Section 3.3.1). Thoughtful presentation can help, and in particular visual interactive presentation may support exploration and understanding of the structures that these models uncover. While topic modeling is used abundantly as a component in various applications, some previous works focus on visualizing topic models specifically, e.g., by structured presentation of text with little other visual encoding of information [75, 45], and some relying more extensively on graphical representation of the relationships of the model [53, 80]. For a recent and thorough comparative study of different topic visualization techniques, including word lists, tag clouds and network visualizations, confer Smith et al. [188]. They observe that network visualization may help users summarize topics in more abstract and descriptive terms.

Two forms of visual topic exploration are presented in this section, both utilizing network visualization as a means to communicate inherent graph structures. In contrast to the co-occurrence network based on syntagmatic relations, these networks extend the notion of co-occurrence and are able to reflect semantic/thematic word associations as well. In the case of LDA, the probability distributions define links that can be visually represented, and in the case of word embeddings, semantic similarity measurement can generate useful links. Interactive force-directed layouting is used in both cases to give a natural impression of the global topology of the thematic structure and of local relationships, as well as to support interpretation by letting the user browse and highlight details on demand.

[4] See https://www.bloomberg.com/view/articles/2016-04-18/the-financial-threats-that-machines-can-see

[5] For instance, regarding the importance of the interpretability it provides, see https://www.centralbanking.com/technology/3270121/teaching-machines-to-do-monetary-policy

Data. As a sample case, a corpus consisting of 3954 patent application abstracts filed between 2001 and 2011 is analyzed. The data stems from the U.S. Patent and Trademark Office (USPTO)[6] and consists of financial/business method patents, filed under patent subclass 705/35 (relating to banking, investment, credit, etc.).

Applications and grants for this type of patent has soared in the U.S. and caused a *patent flood* [139] following a decline of the business method exception to patentability promoted by certain court rulings, as well as a lack of an appropriate classification standard. The increased issuance of low-quality patents, which are defined in vague, or overly broad terms or overlap other patents, complicates the review process and the search of prior art [84]. This has stimulated research into patent information retrieval and patent mining [130, 226], where visualization nevertheless has played a rather minor role so far. Although not customized toward this problem in particular, the topic modeling and visual exploration discussed here use this case to demonstrate their function and it also serves as a basic example of how the patent search problem may be addressed.

4.2.1 Topic model visualization using graphs

The application using LDA topic modeling, presented in Paper III, is summarized in the following. This visualization method interprets the topic model as a graph structure and visualizes it as a network of keyterms and topics. The topic-term probabilities β_k provide the basis for linking terms and topics, and the topic-specific distinctiveness of a term $P(k|w)$, also introduced in Section 3.3.1, provides the link weighting. The most distinguishing terms per topic (with highest $P(k|w)$ per k) are included in the network, where each topic and each unique term is represented as a node. Terms are associated to each other by second-degree connections over the latent topics.

Results. Some keyterms are shared among topics, and they connect and relate topics in an interpretable way. The force-directed layouting places a shared term between its topics, and the relationships are plotted explicitly by lines. These indirect topic-topic links result in a spatialization that communicates a topic similarity structure and gives an impression of the global thematic composition of the corpus, as is illustrated in Figure 4.5. The prevalence of a topic, as measured by its average probability over documents, is represented by node size. This helps to communicate the topic composition more truthfully.

Figure 4.6 shows a focused view on a single topic, where the link weights are encoded by link opaqueness and the distinctiveness of all terms toward

[6]The publicly sourced data set has been prepared by Fredrik Lucander.

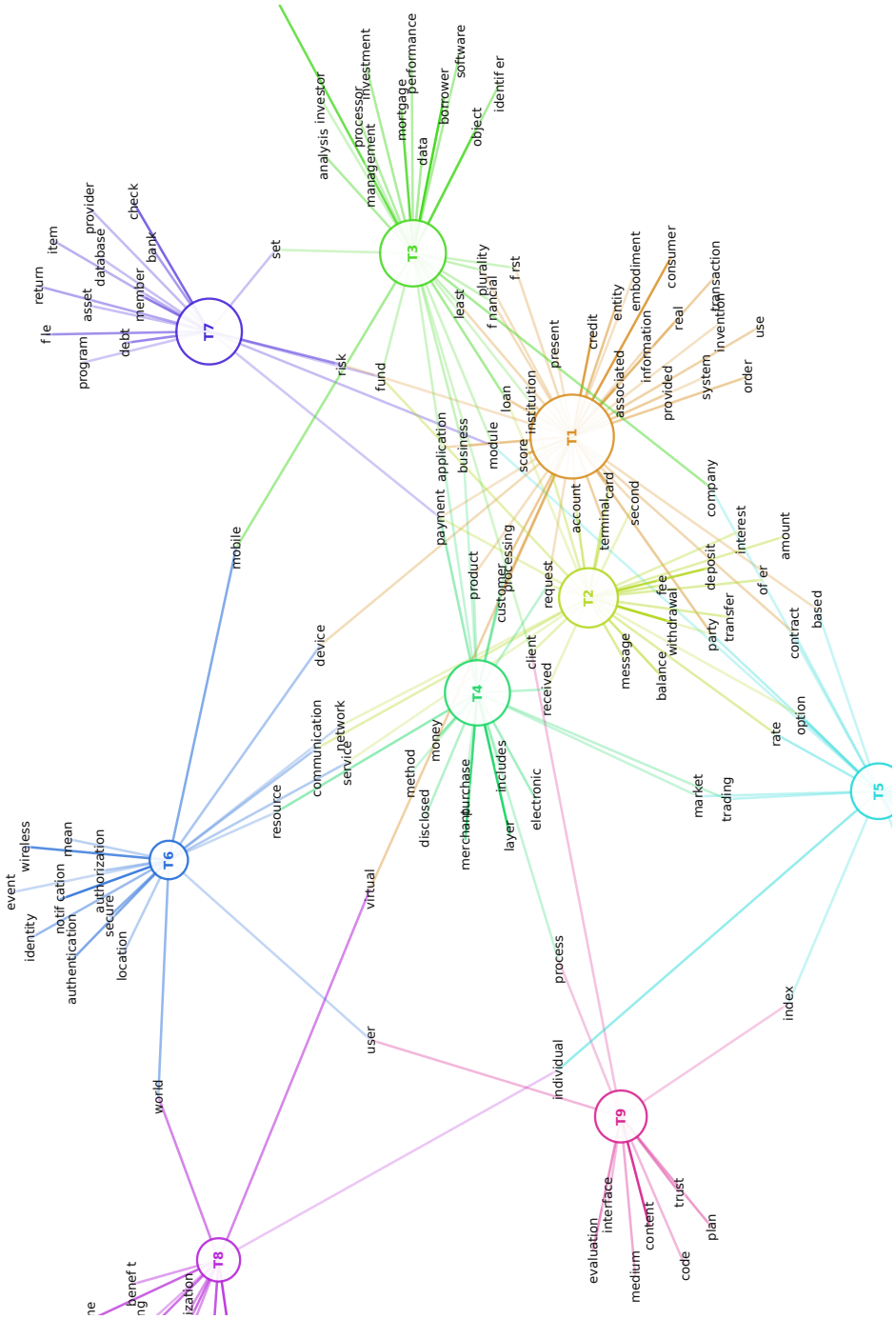


Figure 4.5: Overview (cropped) of LDA topics in financial patent abstracts, illustrating links between topics (circles) and most distinctive keyterms, as well as relatedness between topics in terms of shared keyterms. Interactive demo and source code available at: http://samuel.ronnqvist.fi/topicGraph/

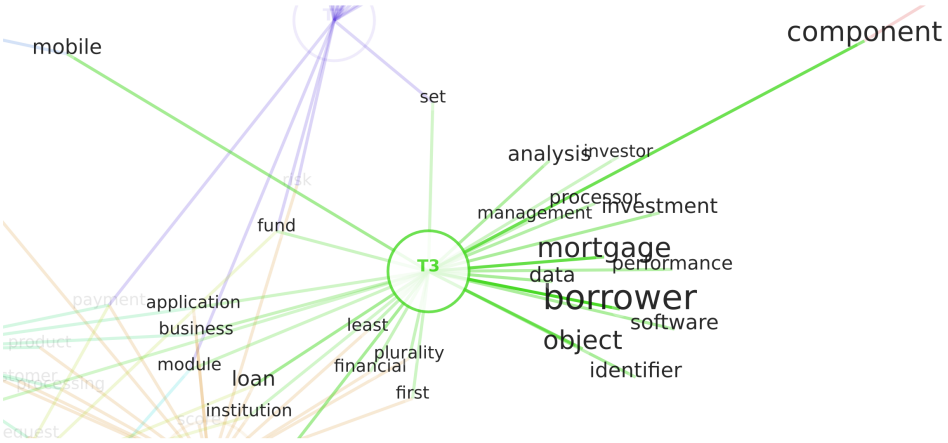


Figure 4.6: Focused view on single topic, representing keyterm distinctiveness by font size for pop-out effects.

that topic are encoded also by font size. Exploiting this sensory representation, in the more distinguishable visual channel of size variation, creates an easy-to-read tag cloud. This view also dims the rest of the network to shift attention onto the topic and keyterm nodes, while the rest remain visible as context. Likewise, the interface can focus on a single term by hovering and highlighting its associated topics, in order to support visual search.

The network supports zooming and panning for navigation, as well as dragging of the nodes to explore different arrangements and to resolve possible overlaps. The network as a whole provides a scaffold for information retrieval, where the selection of single or combinations of terms and/or topics can query the underlying documents. The topic-document probability θ_d and term-frequency per document may link terms and topics to documents, and support their ranking and retrieval.

The figures visualize a topic model with 10 topics extracted from the patent corpus. Figure 4.5 shows that topic T1 is the most frequent and that it is thematically central to the corpus. A few other topics are similar in size and centrality, while others are somewhat smaller and more peripheral, some even outside the figure crop. In the focused view in Figure 4.6, T3 can be interpreted as a topic about loans, while topic T2 seems to represent payment systems patents and T6 patents on authentication systems. Topic T1 appears similar to T3, both discussing credit, where closer inspection can differentiate them further as relating to technological aspects of credit risk management and mortgage services respectively. Finally, shared terms, such as *mobile* between T3 and T6, can highlight the commonality of the linked topics, while the contexts of the respective topics also help to disambiguate the shared term.

The topics T3 and T1 may be compared against the USPTO classification scheme, which defines a subcategory 705/38 ("credit (risk) processing or loan processing (e.g., mortgage)") under subclass 705/35. This hints at the potential value of topic modeling or other knowledge-free approaches as support for more granular and data-driven patent search.

The visualization provides an overview of the topical structure of the patent corpus, and some means to interpret, name and relate topics. A graph-based visualization of topic models may bring forward structures that are otherwise difficult to observe, but a fundamental challenge remains in interpretation of the individual topics. Taken out of their sentence contexts, the keyterms are only interpretable with the support of other extracted keyterms. Reading these topics can be rather demanding and confusing, with possible problems of overly broad, narrow, overlapping or incoherent topics that to some extent can be countered by parameter configuration. Nevertheless, these issues add to the challenge of inferring higher-level meaning to the collection of keyterms. The following section explores whether departing from the notion of distinct topics, and modeling word-level semantics directly can offer a more intuitive way of exploring themes in text.

4.2.2 Topic modeling with word vectors

LDA topic modeling and word embeddings are similar in the sense that they both infer distributed latent distributions of terms, and while the use of LDA tends to focus on understanding topics and organizing documents, word embeddings focus specifically on the relations among words. In the topic modeling context, words are the most directly interpretable and quantifiable units of meaning that we can extract, which prompts the question if not exploration of topics may be more naturally organized around words in presentation, while the latent representations are kept in the background. This section introduces the work of Paper IV on this topic.

Word vectors, as introduced in Section 3.3.2, are based on co-occurrences between a word and its observed context words, and this information is represented in dimension-reduced form in a dense vector. These latent representations are likely difficult to interpret as such, and need not be shown. Instead, words are directly compared based on the similarity of their representations. In this case, cosine similarity is measured between skip-gram vectors, in order to construct a network of terms (the algorithm is listed in Paper IV). In order to study the most prevalent topics of a corpus and reduce the complexity of the network visualization, only the most frequent terms (e.g., top-1000) are compared.

Results. For the purpose of modeling topics, the skip-gram method is applied to considerably smaller data sets, than the billions of words that are

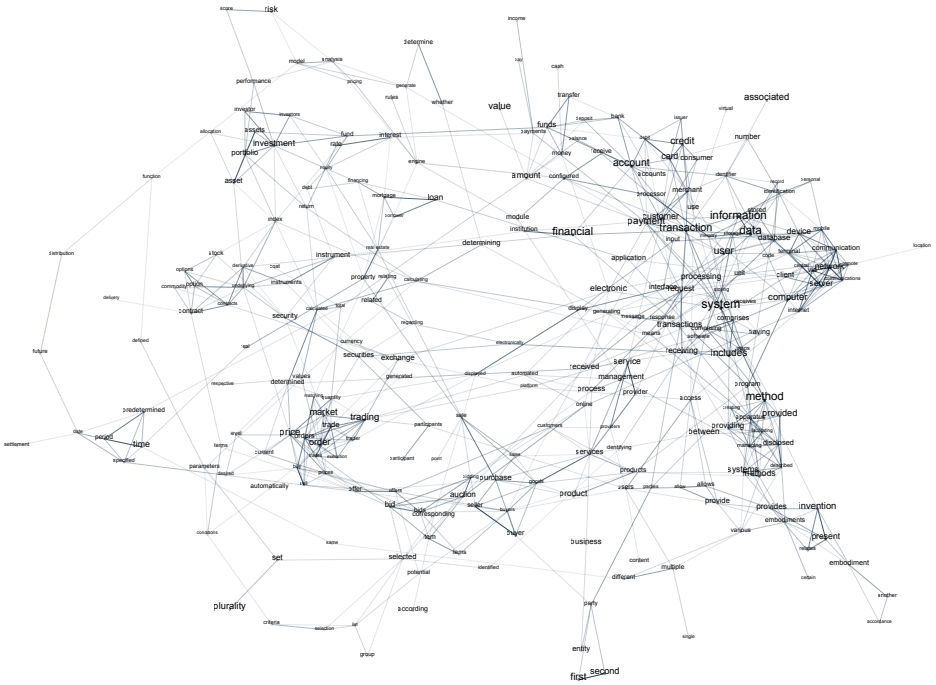


Figure 4.7: Overview of topics in financial patent abstracts, visualized through semantically-organized keyterms. Interactive demo and source code available at: http://samuel.ronnqvist.fi/topicMap/

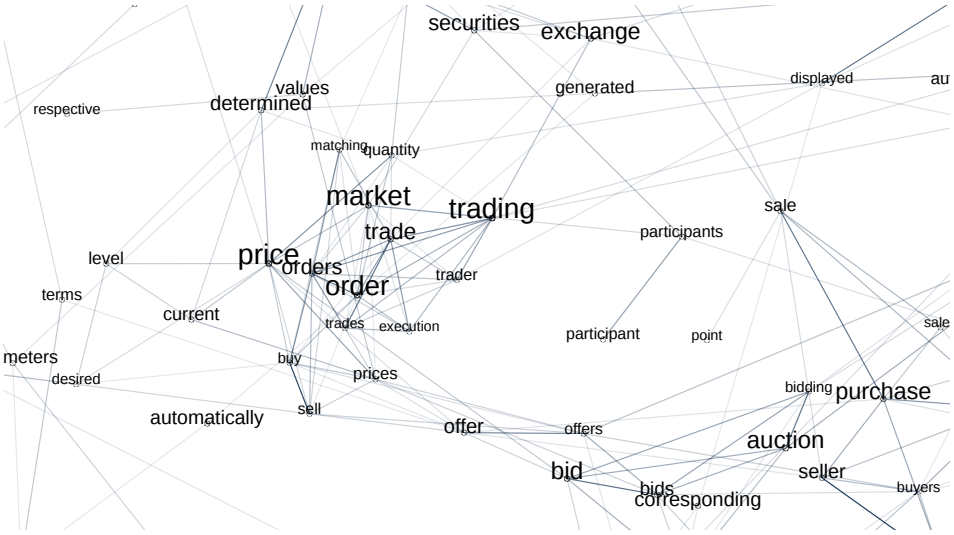


Figure 4.8: Detail of the word-vector-based topic visualization for financial patent abstracts.

often used to produce high-precision embeddings [142], but it is nevertheless able to identify semantic relationships that are useful. The vectors associate paradigmatically related terms based on their similar contexts, and terms that frequently occur in a syntagmatic relationship also tend to register similar contexts and become related. The paradigmatic relations provide a semantically coherent organization of the network, while the syntagmatic relations can serve as disambiguating context to neighboring words.

The network is constructed with a lower bound on the term similarity scores, and an upper bound on number of links per term node, in order to produce a sparser network that can be visualized well. Force-directed layouting is then used to provide an intuitive map of the frequent terms in the corpus, which is semantically organized to support browsing. Term frequency is encoded by font size as in a tag cloud, whereas the semantic spatialization and explicit linking (with similarity-encoding opaqueness) create a coherent view. The interface offers similar interaction as discussed in the last section, e.g., for highlighting of a term's immediate neighborhood, and it supports linking of information to this overview. Topics emerge in the network visualization without explicit borders, as more densely connected regions of related words, as well as gradients across the map of gradually shifting meaning.

Figure 4.7 shows the visualization for the patent corpus, where a main dense region is centered around the very frequent and general terms *system* and *method*.[7] More narrow topics can also be observed, such as the nearby dense cluster of the terms *computer, server, network, communication, device, mobile,* etc. From there, a gradual thematic shift can be observed toward *database, data, information, transaction, payment,* etc. Figure 4.8 shows a close-up view of another cluster with terms such as *price, market, trading* and *order,* which lets one identify trading-related patents as a particular topic. The semantic organization makes the visual search intuitive, and along with the absence of explicit separations it may alleviate the problem of incoherence.

While LDA serves both exploration and representation learning purposes, this approach targets exploratory analysis specifically. Within the framework discussed in Section 3.1, it provides an example of the combination of machine learning and human intelligence, where the computational part can scale well to large corpora and provide meaningful organization for the user, who makes sense of the content through integration with their knowledge about the world and the task. This achieves knowledge-free text analysis, as knowledge does not need to be externalized and encoded, but continuously integrated during the exploration process.

[7] These terms were not visible in the LDA-based network, as they are not distinctive of any topic. The raw topic-term probabilities rank these terms at the top for every topic.

The topic modeling approach to text analysis is very flexible and can serve as a useful tool for exploring corpora, at least at a general level. Yet, as discussed at the end of Section 3.3.2, breaking down text into words or compound words enables their quantification, while it also takes them out of context and makes interpretation more difficult and unreliable. The open question is how to support and pursue knowledge-free text analysis, which can quantify aspects of the text, and thus utilize computational modeling, while also providing more context for interpretation. The solution may need to utilize both machine learning in novel ways and clever interface design.

4.3 Multilingual shallow discourse parsing

Problem. The task of shallow discourse parsing, introduced in Section 3.4.2, is here studied in a multilingual setting. The CoNLL 2016 Shared Task challenge [221] focused on shallow discourse parsing of both English and Chinese, as an extension to the English-only task of the previous year. Hence, abandoning a rich-feature-based approach in favor for representation learning allows systems to be developed that can easily be adapted to new languages, by retraining rather than reengineering of the system or the features.

The shared task involved the recognition of discourse argument spans and connectives, as well as classification of the relation between two arguments. While Paper V describes the *Frankfurt Shallow Discourse Parser*,[8] an extended discourse parsing system that handles other subtasks as well, this section focuses particularly on the module for implicit cases and the classification of relation sense. This is the most challenging problem,[9] and, as a frequent phenomenon, it is important to solve in order to enable reliable parsing of discourse structure in text.

Data. Annotated data is nevertheless needed for the task, and suitable corpora are available for English and Chinese. The shared task uses the Penn Discourse Treebank 2.0 (PDTB)[10] [171] for English and the Chinese Discourse TreeBank 0.5 (CDTB)[11] [228] for Chinese, which follow compatible annotation schemes. Both corpora are based on newswire text.

The PDTB corpus contains 34k relations for training and validation of which 53% are implicit, whereas CDTB contains 11k of which as many as 76% are implicit. This indicates that coherence relations are considerably

[8]Source code available at https://github.com/acoli-repo/shallow-discourse-parser

[9]The best performing systems in the shared task tested 90.22% accuracy on sense classification for English explicit cases, compared to 40.95% for the implicit. For Chinese, the best test accuracies were 96.84% (the system presented in Paper V) vs. 72.42%.

[10]Available at https://catalog.ldc.upenn.edu/LDC2008T05

[11]Available at https://catalog.ldc.upenn.edu/LDC2014T21

Senses for English	Freq. (%)	Senses for Chinese	Freq. (%)
Comparison	0.84	Causation	2.33
Comparison.Concession	1.12	Conditional	0.33
Comparison.Contrast	9.34	**Conjunction**	**66.29**
Contingency	0.01	Contrast	0.83
Contingency.Cause*	19.86	EntRel	14.07
Contingency.Condition	0.01	Expansion	15.2
EntRel	**23.91**	Progression	0.09
Expansion	0.42	Purpose	0.72
Expansion.Alternative*	0.88	Temporal	0.14
Expansion.Conjunction	18.67		
Expansion.Exception	0.01		
Expansion.Instantiation	6.55		
Expansion.Restatement	14.38		
Temporal	0.01		
Temporal.Asynchronous*	3.14		
Temporal.Synchrony	0.88		

Table 4.1: Class distributions for the English and Chinese training sets, implicit relations only (including EntRel). The five third-level classes are joined with their parent classes (*) for compact view.

more often expressed only implicitly in Chinese. The prevalence shows that being able to handle implicit relations is particularly important, and it motivates the focus on recognizing implicit relations and the inclusion of Chinese in the applications described below. Implicit relations in this context encompasses also entity relations (*EntRel*), as they lack a connective, while the explicit set includes alternate lexicalization relations (*AltLex*) accordingly.

Considering only the implicit cases, the English data contains 20 senses (defined in up to three levels, e.g., *Contingency, Contingency.Cause, Contingency.Cause.Reason*), of which 9 occur more frequently that 1% in the training set. *EntRel* is the majority class with a frequency of 23.91%. The Chinese training set contains 9 senses, of which 4 are more frequent than 1%. The *Conjunction* majority class at 66.29% sets the initial baseline for the prediction task considerably higher than for English, which makes the task especially challenging. Table 4.1 summarizes the class distributions of the data used for training, in order to illustrate the modeling problem.

4.3.1 Feed-forward network on English and Chinese

This section presents the application of the first model described in Section 3.4.2 and Paper V, namely the feed-forward neural network on bags of vectors. The parser containing this model participated in the main task of the

challenge that included recognition of arguments and connectives, as well as in a supplementary evaluation that only tested the accuracy of sense classification. The following discussion focuses on the latter form of evaluation for the implicit (non-explicit) cases, as this measures the quality of the output produced by the feed-forward model itself, avoiding error propagation in the parser pipeline.

Results. The model is applied using pre-trained embeddings provided within the scope of the challenge,[12] namely *GoogleNews* word vectors for English and comparable vectors for Chinese trained by skip-gram on the *Gigaword* corpus. The vectors have a dimensionality of 300 in both sets. In order to improve performance of the model, the vectors are tuned in an unsupervised manner toward the news text from which the annotated discourse relations are extracted. This is performed by updating the vectors through skip-gram training in multiple iterations, with a low initial and decreasing learning rate. The procedure adapts the pre-trained vectors to the data set of the task and infers representations for missing tokens, such as punctuation marks, which may be especially important in discourse parsing, and was observed to improve accuracy by 3-4% absolute on the development set.

In addition to the embedding information, the model also makes limited use of syntactic dependency information by means of a scheme that weights each vector $V(j)_i$ according to the depth $d(j)_i$ of token $t(j)_i$ by $\frac{1}{2^d}$, with depth measured from the root node of the sentence. Evaluated on the development set, this depth-weighting scheme improved performance by about 1.5%.

The model and the above mentioned techniques were developed against the English data. In order to apply it to the Chinese data, the hyperparameters were reoptimized but otherwise the model was kept intact. This demonstrates the model's ability to adapt to new languages. The optimal networks consist of 600 input nodes and 40-60 hidden nodes, trained with momentum and regularized by L1 or L2.

The official evaluation results[13] are summarized in Figure 4.9, focusing on performance on the test set, which stems from the same source but was held out from model training and validation. For English implicit (non-explicit) sense classification, the model scored 37.61% accuracy (rank 4) compared to the best system reaching 40.91%. For Chinese, the model scored 71.87% (rank 2) against the leading 72.42% on implicit senses, and helped achieved state-of-the-art accuracy on overall sense accuracy (implicit and explicit). The results show that the model and the parser as a whole are

[12] See http://www.cs.brandeis.edu/~clp/conll16st/dataset.html

[13] See http://www.cs.brandeis.edu/~clp/conll16st/results.html

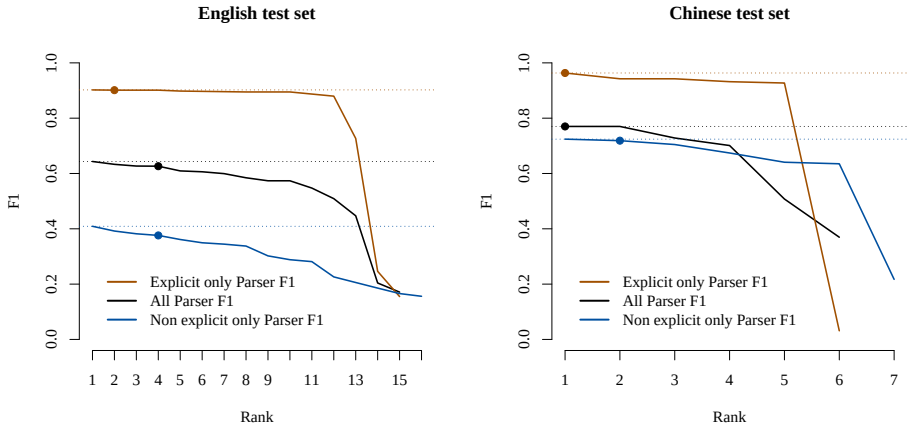


Figure 4.9: Official CoNLL 2016 Shared Task evaluation results for sense classification, with the Frankfurt Shallow Discourse Parser (circle) related against all other systems. The reported F1-score is in this case equivalent to accuracy.

competitive, and that this level of performance is achievable without the use of rich linguistic features. The model even ranks higher on Chinese, which is remarkable considering that its structure was developed against English data only. The parser classifies explicit relations using a linear-kernel Support Vector Machine with the connective as the only feature (in accordance with Chiarcos & Schenk [51]), a minimalist approach that given the recognition of connectives and the discourse relation training data does not require any further resources. The full parser has been made open source.[8]

4.3.2 Attention-based recurrent network on Chinese

The second model for implicit shallow discourse parsing, introduced in the latter part of Section 3.4.2 and described in detail in Paper VI, extends the previous by introducing recurrence for word-order-aware modeling. The Attention-based Bidirectional Long Short-Term Memory (Att-BLSTM) network utilizes an attention mechanism for dynamic weighting over the sequence in the classification process, which provides insight into what drives the model's decision.

Results. The Att-BLSTM model is linguistically resource-lean, utilizing only pre-trained word embeddings (Chinese Gigaword). While developed on the CDTB data set, it is language independent by design. The application toward Chinese implicit sense classification is motivated by the relative

scarcity of tools for Chinese compared to English, and by the aforementioned prevalence of implicit relations in Chinese discourse. The model is evaluated by the framework established through the CoNLL 2016 Shared Task, which enables reliable comparison among systems based on consistent setups and data.

The model is trained on sequences of 256 tokens with leading zero padding or truncation of the original sequences. Hence, the 5-layered model has a depth of up to 5+256, in terms of Credit Assignment Path (cf. [184]). The dimensionalities of the embedding layer and the LSTM hidden vectors are all 300. Dropout regularization at a 0.5 rate is applied between layers from embedding to output. LSTM inputs also are L2-regularized, whereas recurrent connections are fully unregularized to avoid detrimental loss of information over longer distances in the sequence (cf. [224]).

A partial argument sampling scheme is introduced to improve performance, which extends the training and development data sets. Each data point consisting of an argument pair and a sense label (a_1, a_2, y) produces the set of data points $\{(a_1, a_2, y), (a_1, a_2, y), (a_1, y), (a_2, y)\}$, which includes single-argument instances as well. This procedure is motivated from a representation learning perspective, as it may stimulate learning of representations for individual discourse arguments, which then supports their compositional representation and modeling. This line of reasoning is followed also by LeCun et al., writing: "these advantages [of deep networks with distributed representations] arise from the power of composition and depend on the underlying data-generating distribution having an appropriate componential structure." [122] The recurrent model is assumed to exploit this power of composition going from pre-trained word-level representations to representations of the sequence, and not the least there is componential structure of the argument pair that the model should capture. The partial sampling is also motivated from a linguistic point of view, as previous work has shown that isolated arguments can evoke strong expectations of implicit discourse relations [9, 176], which further suggests a componential structure underlying discourse relations, and that encouraging representations of single arguments to be learned may support modeling of the joint structure. The effect of the sampling scheme is evaluated on the test set, yielding an absolute increase of 5.74% accuracy over the original data.

The recurrent modeling, evaluated on the CDTB test set, is able to achieve state-of-the-art accuracy of 73.01%. This is achieved using the attention mechanism, which yields an absolute increase of 2.70% against using only the final LSTM hidden vectors. Many of the comparable systems utilize feed-forward neural networks, including the previous best system of Wang and Lan [208]. In contrast to assertions by Rutherford et al. [181], this shows that recurrent networks are able to outperform feed-forward networks on implicit sense classification, although recurrent models may be more dif-

Figure 4.10: Visualization of attention weights over discourse arguments. Darker blue cells represent tokens that receive more attention by the model as it classifies the relation sense (conjunction, entity relation).

ficult to train reliably and require considerably more computational power for training. Feed-forward architectures seem to remain a viable lightweight alternative.

Finally, as the sense classifier has been trained and evaluated, Figure 4.10 illustrates its function by two examples drawn from the CDTB test set. The arguments of the example discourse relations are shown in Chinese with English translations underneath. The model has correctly classified the most likely senses for the two relations: *Conjunction* and *EntRel*. As the network reads the sequence containing the argument pair, the attention layer produces a weighting dynamically based on the input. The weighting is visualized token by token and describes to which parts of the sequence the model attends as it makes its decision. The colored cells encode the level of attention by intensity, darker blue being assigned higher importance.

In the first example, the model focuses around the argument border, where the semantically related terms *understandings* and *agree* occur, as it identifies the relations as *Conjunction*. In the second example, the entity relation (*EntRel)* is identifiable by the references from the second argument (*Governor*, *Macao question*) to the entity in the first (*Macao government*), and attention is placed especially on the second argument where the references are found and hence the relation recognized.

The attention visualizations offer some transparency into the functioning of the otherwise opaque deep learning model, and may thereby provide some insight into the linguistic phenomenon.[14] Together with other similar efforts (see, e.g., [220, 128, 49]), the visualizations sketch a picture of how the success of deep learning in natural language processing and other artificial intelligence applications (see Section 2.2.3) can be exploited, while seeking to open up the models and make them more interpretable. This is an important focus, recognized by efforts such as the Distill Research Journal,[15] as a lack of transparency and understanding of advanced machine learning models may create skepticism, and hold their dissemination and successful application back in many areas.

The Attention-based Bidirectional Long Short-Term Memory network has demonstrated state-of-the-art performance on the task of classifying Chinese implicit discourse relations, and the work furthers the understanding of how recurrent modeling can serve discourse parsing. The attention mechanism both boosts the predictive performance and offers transparency of the deep model. In order to encourage further development on this particular problem, the model is open sourced as the first of its kind.[16]

[14]Furthermore, attention visualization based on bar charts that highlight patterns across multiple instances of the two relation types have been introduced by Niko Schenk [182], as a means of characterizing the differences between the senses from a linguistic point of view. A visualization to this end was jointly developed and presented at the ACL 2017 conference, and is available at: https://github.com/sronnqvist/discourse-ablstm/blob/master/acl_poster.pdf

[15]See http://distill.pub/journal/

[16]Code available at: https://github.com/sronnqvist/discourse-ablstm

Chapter 5

Conclusions

This thesis has been positioned within a data mining context, introduced from the perspective of navigating an information-rich environment; i.e., the aim of data mining is to make sense of a world represented through abundant data. In the introduction, I related the urge to analyze such data to the essence of intelligence, as an instinct to seek meaning in everything we perceive, and described the use of computational methods as a natural way of extending one's capabilities.

Data mining is often motivated in the literature by a perceived information overload, but it may be naive to think that the ability to keep pace with information is a unique concern in the digital age. Already in 1755, Diderot projected that the growth in information would exceed the human capacity for processing, as he wrote:

> "As long as the centuries continue to unfold, the number of books will grow continually, and one can predict that a time will come when it will be almost as difficult to learn anything from books as from the direct study of the whole universe. It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes. When that time comes, a project, until then neglected because the need for it was not felt, will have to be undertaken." [62, p. 85]

His words seem to describe what we know as text mining. While he, in the spirit of the Enlightenment, may have been driven more by the curiosity of scientific discovery, many projects today apply computational analysis for the sake of providing a competitive advantage. Irrespective of the reason,

there is a widespread demand for the ability to make use of data, including text. However, there is at the same time a shortage on experts to meet this demand.

I have motivated the knowledge-lean approach to text mining by its focus on minimizing the need for certain expert knowledge, with the goal of making computational text analysis more widely applicable and allowing it to find previously unimagined uses. In particular, lessening the need for linguistic expertise may unlock application in domains that would not be able to support such interdisciplinary work. As such, my intention has been to encourage the spread of text mining, allowing it to find its way into a plurality of many less frequented areas of application, where it may serve interesting and beneficial causes. I consider this a process to democratize text mining.

As regards understanding the information around us, we may always be wishing for more and better tools. Our aspiration to comprehend may be relative to the level of information available. Nevertheless, observing the increase in data within the last few decades, which has been tremendous in absolute terms, might add a sense of urgency. Much of the data is generated in centralized locations, and therefore access becomes unevenly distributed. I consider it worthwhile working to even the competition, which might be achieved by placing the ability to use the tools in the hands of the many. Furthermore, as I have shown regarding the data access problem in systemic risk analytics, tools can be designed with limited means to exploit openly available data sources.

In the following, the work presented in this thesis is summarized, and some current limitations are discussed alongside ideas for future research directions.

5.1 Summary

The thesis at hand has focused on how the design of text mining methods can support practical analysis needs in a knowledge-lean manner. An application-oriented pragmatism has called for the incorporation of some encoded knowledge, while the challenge has been to keep the requirements at a minimum. The rationale has been that intensive use of knowledge resources, that may not generalize well across tasks, domains or languages, may restrict the use of text mining to well-defined and well-funded problems, as well as to well-supported languages. Other peripheral-interest application areas or undersupported languages may not be able to benefit from many traditional, knowledge-intensive forms of text mining. This thesis has studied flexible, data-driven and knowledge-lean approaches that may support early exploration in such new text mining territory. The methods may serve to scout

the landscape for more targeted and more knowledge-intensive use of text mining that may follow.

The work has combined computational modeling, most importantly machine learning and the learning of semantic representations, with visualization, with an aim to leverage the strengths of the respective forms of data analysis: computational and human. On the one hand, knowledge-lean text mining is made possible by unsupervised modeling that is able to extract meaningful structure from text on its own. On the other hand, human exploration and assessment of model output, which seamlessly incorporates knowledge into the process without the need to encode it, is an integral part of the process.

The methods presented include modeling of direct relations expressed in text and indirect relations as a basis for semantic association. The identified structures are visualized to enable exploring and understanding the detailed information, or used as data representations for predictive modeling. Predictive modeling, in turn, serves to further structure and simplify text for human analysis. The visual-analytics-based framework offers a way to intimately combine human and machine intelligence in a joint text analysis process.

The thesis has presented the application of these methods to concrete problems. First, within the field of systemic risk analytics, the applications have explored the utility of text as a novel source of information in monitoring financial stability by computational modeling. This area constituted a suitable case for testing knowledge-lean text mining in a new domain of application. The work has received recognition within the domain and set foundations for further text-based efforts within systemic risk analytics. Second, the thesis has studied modeling and visualization of topics in corpora, and tested the methods on patent texts for data-driven discovery of thematic structure. Third, the knowledge-lean approach has been applied for natural language processing, namely discourse parsing that does not rely on rich linguistic features, which may support knowledge-free forms of text mining by bringing forward more refined structure. The work has shown that linguistically resource-lean modeling of the task is able to achieve competitive results, with one of the introduced models achieving state-of-the-art performance for Chinese.

The applications have focused on practical text analysis needs, and the thesis has presented the underlying methods that serve to structure text in various ways for analysis. The problem of costly and highly specialized linguistic resources has been mitigated by starting with open-ended and highly knowledge-lean types of analysis. Targeting the analysis toward event prediction was achieved in a knowledge-lean manner by decoupling the supervision data from linguistic form. The introduction of discourse parsing, as a method for refining retrieved event descriptions, demonstrated a trade-

off between sophistication of language processing and knowledge leanness. Even as it was achieved in a resource-lean manner relying on representation learning, the processing required language-specific annotated data.

5.2 Limitations and future research

The thesis has tied together work based on a few in-depth studies, in order to sketch an overall picture of how knowledge-lean text mining can be pursued. Next, knowledge-lean text mining is discussed at a general level, followed by commentary relating to the particular application areas of the thesis.

A natural development of the presented work would be to pragmatically extend into particular domains and focus on more specific problems. While this may be possible, continuing to adhere to a knowledge-lean philosophy, it is set to gradually become more domain-specific and likely also gradually more knowledge intensive. This, I argue, illustrates the suitable place for knowledge-free and knowledge-lean text mining, as a tool for early exploration within a particular area. As the analysis objective becomes more defined, the approach may give way to more knowledge-intensive text mining that can excel at specific problems. Yet, it will benefit from data-driven methods that improve the efficiency of knowledge resources.

Fully knowledge-free text mining does have some inherent limitations, as has been discussed throughout the thesis. While text contains meaningful structure that can be discovered by unsupervised means, encoded knowledge, if available, generally offers a very practical shortcut, and sometimes even irreplaceable information. Navigating around the need for encoded knowledge has been the central theme, and often a pragmatic compromise is the best way forward, even if that means departing from knowledge free into the knowledge lean. The discussion regarding discourse parsing has also highlighted some presumed outer limits of truly knowledge-lean text mining, where annotation-based modeling of language is accepted as the cost for improving the utility of the text mining system. Likewise, the distant supervision in modeling of events meant a departure from the completely knowledge free, albeit with a rather lightweight requirement on domain-specific information not tied to linguistic form.

I imagine that there is a core of purely knowledge-free text mining, such as topic modeling and other fully unsupervised approaches, whose value lies in their open-ended nature where the user is free to integrate observations with held knowledge. Moving away from that core, there are various directions in which to make the analysis more targeted. This may be done interactively, e.g., focusing on a specific word in a topic model visualization. More often, however, the targets are specified and necessary knowledge embedded prior to modeling. Consequently, layers of knowledge encoding of

various kinds form around the core. This thesis has sought to characterize the role of knowledge leanness in text mining and chart its core and inner layers, through a general discussion and through a few examples of pragmatic domain-motivated endeavors into somewhat more knowledge-intensive spheres.

Beyond this thesis, the topic of knowledge-lean text mining is little explored as such. Continued and more thorough study of its characteristics and potential extent would therefore be interesting. Nevertheless, the most practical value is likely to be had from combining data-driven and unsupervised modeling with knowledge-based approaches as needed, where this work may offer guidance.

Systemic risk analytics. Finally, a few projections for continued work on the application areas should be plotted. The use of text data in systemic risk analytics is still in its infancy, and further research is needed on how to integrate text data in meaningful ways from a domain perspective. Maturing the focus of such applications takes time and requires good understanding both of the domain with its needs and of the text mining methods. The interdisciplinary work is challenging and requires a dissemination of knowledge in both directions.

Concretely, next-step extensions of the presented work may involve the combination of semantic modeling and co-occurrence network analysis, for constructing better-defined and interpretable network models. While the discussion regarding event modeling has focused here mainly on an information retrieval and qualitative focus of describing events, pure quantitative performance and forward-looking capabilities are particularly compelling to economists. In particular, integrating traditional data sources with text-based modeling should provide better predictive performance and fulfill the goal of having text data serve as a complementary source of information. The text processing itself has plenty of room for improvement as well, e.g., refined text-based network models may capture the bank interconnectivity dimension for predictive modeling. Moreover, a focus on forecasting distress rather than focusing on coinciding events, while more challenging to identify relevant signals in text, stands to offer considerable value. A worthwhile direction to explore this would be based on more opinionated, analytical and forward-looking text material, where semantic analysis may focus on picking up relevant sentiment of experts or miscellaneous crowds to be treated as proxies in the modeling of financial risk.

Topic modeling. Due to the knowledge-free nature of topic modeling, it remains an approachable and popular tool in many areas where both linguistic resources and text mining skills are scarce, e.g., within digital

humanities, sociology and political science. This alludes to the value of its general approach, i.e., the knowledge-free modeling that supports open-ended analysis. Thus, a question remains: how could the general idea of topic modeling be made more useful without compromising its broad appeal?

Making it more useful might mean at least two things. First, it means making the analysis more targeted. How is this achieved in a manner general enough not to require technical skill and exclude users? Can this be specified in a user-friendly way, and how does it affect modeling requirements? Second, it means making the model output easier to interpret. As has been discussed, an inherent limitation of topic modeling is that by breaking down text into words or expressions context is lost, making interpretation more difficult. Along this line, future work should explore ways to improve interpretability, by providing more context directly or through interactive exploration, while seeking to retain general applicability and ease of use.

Discourse parsing. In this thesis, discourse parsing was positioned as a tool for introducing structure to knowledge-lean text mining, with the aim of improving interpretability of its output. Other forms of natural language processing may be well posed to work toward that same aim as well, and more work should investigate how to embed discourse parsing in text mining applications in practice. Evaluation based on higher-level applications should constitute a meaningful direction of development for discourse parsing.

The CoNLL Shared Task challenges of 2015 and 2016 laid important groundwork for shallow discourse parsing, by establishing a common framework for evaluation and gathering research interest around the problem. This will likely continue to yield efforts of improved modeling by innovative use of machine learning, which may incorporate more (representations of) real-world knowledge and reasoning based on it. Implicit discourse parsing may be an *AI-hard* problem that requires the ability to represent and work with extensive real-world and common-sense knowledge to reach very high accuracy.

From a knowledge-intensiveness point of view, a problem remains in the need for annotation of discourse corpora. Shallow discourse parsing may be of a somewhat peripheral interest within natural language processing, which is reflected by the lack in training data for languages beyond English and Chinese. At this time, there does not seem to be a viable alternative to the implicit encoding of linguistic knowledge through annotation. Nevertheless, with more work applying discourse parsing in practice, a clearer picture of the important aspects and useful forms of discourse parsing should emerge. Then, it could mature from the current rather theoretical focus centered around the official data sets toward a more pragmatic focus of supporting higher-level applications.

Bibliography

[1] Akerkar, R. and Sajja, P. (2010). *Knowledge-based systems*. Jones & Bartlett Publishers.

[2] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

[3] Ananiadou, S. and McNaught, J. (2006). *Text mining for biology and biomedicine*. Artech House.

[4] Ananiadou, S., Pyysalo, S., Tsujii, J., and Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

[5] Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

[6] Andreeva, T. and Kianto, A. (2011). Knowledge processes, knowledge-intensity and innovation: a moderated mediation analysis. *Journal of Knowledge Management*, 15(6):1016–1034.

[7] Anon (2016). Oxford dictionaries (living english). Accessed 21 November 2016: https://en.oxforddictionaries.com/definition/analytics.

[8] Ashby, W. (1956). *An Introduction to Cybernetics*. Chapman and Hall.

[9] Asr, F. T. and Demberg, V. (2015). Uniform information density at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*.

[10] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[11] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 7.

[12] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

[13] Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247.

[14] Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752.

[15] Bayes, T. and Price, R. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53:370–418.

[16] Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2014). The state of the art in visualizing dynamic graphs. In *Proceedings of the European Conference on Visualization (EuroVis)*, volume 2.

[17] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

[18] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

[19] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

[20] Benton, A., Arora, R., and Dredze, M. (2016). Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 14–19.

[21] Bertin, J. (1983). *Semiology of Graphics*. The University of Wisconsin Press, WI.

[22] Betz, F., Oprică, S., Peltonen, T., and Sarlin, P. (2014). Predicting distress in European banks. *Journal of Banking & Finance*, 45:225–241.

[23] Bholat, D., Hansen, S., Santos, P., and Schonhardt-Bailey, C. (2015). Text mining for central banks. In *Centre for Central Banking Studies Handbook*, volume 33. Bank of England.

[24] Biemann, C. (2007). *Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm*. PhD thesis, University of Leipzig.

[25] Biemann, C. (2012). *Structure Discovery in Natural Language*. Theory and Applications of Natural Language Processing. Springer.

[26] Biemann, C. and Riedl, M. (2013). Text: Now in 2d! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

[27] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[28] Bisias, D., Flood, M., Lo, A. W., and Valavanis, S. (2012). A survey of systemic risk analytics. *Annual Review of Financial Economics*, 4(1):255–296.

[29] Björkelund, A., Hafdell, L., and Nugues, P. (2009). Multilingual semantic role labeling. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL) Shared Task*, pages 43–48.

[30] Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T. (2009). Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP'09 Shared Task on Event Extraction*, pages 10–18.

[31] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

[32] Blei, D. M. and Lafferty, J. D. (2009). Topic models. In Srivastava, A. and Sahami, M., editors, *Text mining: classification, clustering, and applications*, chapter 4, pages 71–94. Chapman & Hall/CRC Press.

[33] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

[34] Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., and Hajič, J. (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

[35] Borsje, J., Hogenboom, F., and Frasincar, F. (2010). Semi-automatic financial events discovery based on lexico-semantic patterns. *International Journal of Web Engineering and Technology*, 6(2):115–140.

[36] Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3: Data-driven documents. *IEEE Transactions on Visualization & Computer Graphics (Proc. InfoVis)*.

[37] Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in statistics*, 1:201–236.

[38] Brandes, U. and Fleischer, D. (2005). Centrality measures based on current flow. In *STACS 2005*, volume 3404 of *LNCS*, pages 533–544. Springer.

[39] Braud, C. and Denis, P. (2015). Comparing word representations for implicit discourse relation classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[40] Cambria, E., Poria, S., Bajpai, R., and Schuller, B. (2016). Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*.

[41] Capet, P., Delavallade, T., Nakamura, T., Sandor, A., Tarsitano, C., and Voyatzi, S. (2008). A risk assessment system with automatic extraction of event types. In *International Conference on Intelligent Information Processing*, pages 220–229. Springer.

[42] Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

[43] Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

[44] Cerutti, E., Claessens, S., and McGuire, P. (2012). Systemic risk in global banking: what can available data tell us and what more data are needed? *IMF Working Papers*, 11(222).

[45] Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[46] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In

Advances in Neural Information Processing Systems (NIPS), pages 288–296.

[47] Chen, D. and Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[48] Chen, H., Cummins, J. D., Viswanathan, K. S., and Weiss, M. A. (2014). Systemic risk and the interconnectedness between banks and insurers: An econometric analysis. *Journal of Risk and Insurance*, 81(3):623–652.

[49] Chen, J., Zhang, Q., Liu, P., Qiu, X., and Huang, X. (2016). Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Assoication for Computational Linguistics (ACL)*.

[50] Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

[51] Chiarcos, C. and Schenk, N. (2015). A minimalist approach to shallow discourse parsing and implicit relation recognition. In *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL) Shared Task*, pages 42–49.

[52] Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

[53] Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM.

[54] Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

[55] Cole, R. and Gunther, J. (1998). Predicting bank failures: A comparison of on- and off-site monitoring systems. *Journal of Financial Services Research*, 13:103–117.

[56] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

[57] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

[58] Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

[59] de Saussure, F. (1959). *Course in General Linguistics*. Philosophical Library.

[60] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391.

[61] Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

[62] Diderot, D. (1987). *The Old Regime and the French Revolution*, chapter The Definition of an Encyclopedia, pages 71–89. University of Chicago Press. Keith Michael Baker (ed.), translation of "Encyclopédie," Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, vol. 5. Paris, 1755.

[63] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

[64] Dreyfus, H. L. (1994). *What Computers Still Can't Do: A Critique of Artificial Reason*. The MIT Press, fourth printing edition. Rev. ed. of: What computers can't do, 1979.

[65] Dreyfus, H. L. and Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. Blackwell.

[66] Eliot, T. (1952). *The Complete Poems and Plays 1909-1950*. Harcourt Brace & Company, New York.

[67] Engelbart, D. (1962). Augmenting human intellect: A conceptual framework. Summary Report AFOSR-3233, Stanford Research Institute.

[68] Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

[69] Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. PhD thesis, University of Stuttgart.

[70] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 363–370.

[71] Firth, J. (1957). *Studies in Linguistic Analysis*, chapter A Synopsis of Linguistic Theory 1930-1955, pages 1–32. The Philological Society, Oxford.

[72] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

[73] Foster, J. B. and Magdoff, F. (2009). *The great financial crisis: Causes and consequences*. NYU Press.

[74] Ganesh, M., Han, E., Kumar, V., Shekhar, S., and Srivastava, J. (1996). Visual data mining: Framework and algorithm development. Technical report, Department of Computing and Information Sciences, University of Minnesota.

[75] Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., and Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. In *Advances in Neural Information Processing Systems (NIPS) Workshop on Challenges of Data Visualization*.

[76] Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hambvrgi svmtibvs F. Perthes & I. H. Besser.

[77] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

[78] Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

[79] Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., and Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

[80] Gretarsson, B., O'donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., and Smyth, P. (2012). Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):23.

[81] Grimes, S. (2014). Naming & classifying: Text analysis vs. text analytics. http://www.huffingtonpost.com/seth-grimes/naming-classifying-text-a_b_4556621.html (Accessed: 18 November 2016).

[82] Grishman, R. (1986). *Computational linguistics: An introduction*. Cambridge University Press.

[83] Gropp, R., Vesala, J., and Vulpes, G. (2006). Equity and bond market signals as leading indicators of bank fragility. *Journal of Money, Credit and Banking*, 38(2):399–428.

[84] Hall, B. H. (2009). Business and financial method patents, innovation, and policy. *Scottish Journal of Political Economy*, 56(4):443–473.

[85] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

[86] Hassel, M. (2007). *Resource Lean and Portable Automatic Text Summarization*. PhD thesis, KTH Royal Institute of Technology, School of Computer Science and Communication (CSC), Numerical Analysis and Computer Science, NADA.

[87] Hawkins, J. and Blakeslee, S. (2007). *On intelligence*. Macmillan.

[88] Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3–10.

[89] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1693–1701.

[90] Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensvie Scientific Discovery*. Microsoft Research.

[91] H.G.Widdowson (1996). *Linguistics*. Oxford University Press.

[92] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

[93] Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eigth Annual Conference of the Cognitive Science Society*, pages 1–12. Amherst.

[94] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

[95] Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., and Nagata, M. (2013). Single-document summarization as a tree knapsack problem. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1515–1520.

[96] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

[97] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

[98] Hogenboom, F. (2014). *Automated Detection of Financial Events in News Text*. PhD thesis, Erasmus Research Institute of Management. No. EPS-2014-326-LIS. ERIM Ph.D. Series Research in Management.

[99] Hogenboom, F., de Winter, M., Frasincar, F., and Kaymak, U. (2015). A news event-driven approach for the historical value at risk method. *Expert Systems with Applications*, 42(10):4667–4675.

[100] Hokkanen, J., Jacobson, T., Skingsley, C., and Tibblin, M. (2015). The Riksbank's future information supply in light of Big Data. In *Economic Commentaries*, volume 17. Sveriges Riksbank.

[101] Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology*, volume 20, pages 87–112.

[102] Huang, H.-H. and Chen, H.-H. (2011). Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

[103] Ieumwananonthachai, A. and Wah, B. W. (1996). Statistical generalization of performcne-related heuristics for knowledge-lean applications. *International Journal on Artificial Intelligence Tools*, 05(01n02):61–79.

[104] Irsoy, O. and Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems*, pages 2096–2104.

[105] Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

[106] Jackson, M. O. and Rogers, B. W. (2007). Meeting strangers and friends of friends: How random are social networks? *The American Economic Review*, pages 890–915.

[107] Ji, Y., Haffari, G., and Eisenstein, J. (2016). A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 332–342.

[108] Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.

[109] Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15.

[110] Karpathy, A., Johnson, J., and Fei-Fei, L. (2016). Visualizing and understanding recurrent networks. In *Proceedings of the International Conference on Learning Representations*.

[111] Keim, D. A., Kohlhammer, J., Ellis, G., and Mansmann, F. (2010). *Mastering The Information Age – Solving Problems with Visual Analytics*. Eurographics.

[112] Keim, D. A. and Kriegel, H.-P. (1996). Visualization techniques for mining large databases: A comparison. *IEEE Transactions on knowledge and data engineering*, 8(6):923–938.

[113] Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3294–3302.

[114] Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

[115] Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, 21(1):1–6.

[116] Kriesi, H. (2012). The political consequences of the financial and economic crisis in europe: Electoral punishment and popular protest. *Swiss Political Science Review*, 18(4):518–522.

[117] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 1090–1098.

[118] Laeven, L. and Valencia, F. (2010). Resolution of banking crises: The good, the bad, and the ugly. *IMF Working Papers*, 146(10).

[119] Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

[120] Lascarides, A. and Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

[121] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

[122] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[123] Legendre, A. M. (1805). Nouvelles méthodes pour la détermination des orbites des cometes. *F. Didot*.

[124] Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

[125] Lewis, R. L. (2001). Cognitive theory, SOAR. In *International Encyclopedia of the Social and Behavioural Sciences*, pages 2178–2183. Pergamon.

[126] Licklider, J. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1:4–11.

[127] Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. Master's thesis, University of Helsinki.

[128] Liu, Y. and Li, S. (2016). Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1224–1233.

[129] Liu, Y., Li, S., Zhang, X., and Sui, Z. (2016). Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2750–2756.

[130] Lupu, M., Mayer, K., Kando, N., and Trippe, A. J. (2011). *Current challenges in patent information retrieval*. Springer.

[131] Magnusson, C., Arppe, A., Eklund, T., Back, B., Vanharanta, H., and Visa, A. (2005). The language of quarterly reports as an indicator of change in the company's financial status. *Information and Management*, 42(4):561–574.

[132] Mahesh, K. and Nirenburg, S. (1997). *The Computer Science and Engineering Handbook*, chapter Knowledge-based systems for natural language processing. CRC Press.

[133] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

[134] Männasoo, K. and Mayes, D. (2009). Explaining bank distress in Eastern European transition economies. *Journal of Banking & Finance*, 33:244–253.

[135] Manning, C. D. and Schütze, H. (2001). *Foundations of Statistical Natural Language Processing*. The MIT Press.

[136] Matthews, P. (1991). *Morphology*. Cambridge University Press.

[137] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[138] McCusker, J. P., McGuinness, D. L., Erickson, J. S., and Chastain, K. (2016). What is a knowledge graph? https://www.authorea.com/users/6341/articles/107281-what-is-a-knowledge-graph/ (Accessed 31 March 2017).

[139] Meurer, M. J. (2002). Business method patents and patent floods. *Washington University Journal of Law & Policy*, 8:309.

[140] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411.

[141] Mikheev, A. (1999). A knowledge-free method for capitalized word disambiguation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, pages 159–166.

[142] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

[143] Mikolov, T., Deoras, A., Kombrink, S., Burget, L., and Cernocky, J. (2011). Empirical evaluation and combination of advanced language moddeling techniques. In *INTERSPEECH'2011*.

[144] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their

compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

[145] Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

[146] Milne, A. (2014). Distance to default and the financial crisis. *Journal of Financial Stability*, 12:26–36.

[147] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272.

[148] Minsky, M. (1952). A neural-analogue calculator based upon a probability model of reinforcement. Technical report, Harvard University Pychological Laboratories.

[149] Minsky, M. (1954). *Neural nets and the brain model problem*. Unpublished doctoral dissertation, Princeton University.

[150] Minsky, M. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press.

[151] Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 236–244.

[152] Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T. (2002). Mining product reputations on the web. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 341–349.

[153] Munzner, T. (2014). *Visualization analysis and design*. CRC Press.

[154] Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press.

[155] Newman, D., Bonilla, E. V., and Buntine, W. (2011). Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 496–504.

[156] Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132.

[157] Nilsson, N. J. (2010). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University Press.

[158] Nonaka, I. and Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

[159] Nyman, R., Gregory, D., Kapadia, K., Ormerod, P., Tuckett, D., and Smith, R. (2015). News and narratives in financial systems: exploiting big data for systemic risk assessment. Bank of England, mimeo.

[160] Okita, T., Wang, L., and Liu, Q. (2015). The DCU discourse parser: A sense classification task. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL) Shared Task*, pages 71–77.

[161] Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251.

[162] Özgür, A., Cetin, B., and Bingol, H. (2008). Co-occurrence network of Reuters news. *International Journal of Modern Physics C*, 19(05):689–702.

[163] Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

[164] Palmer, F. (1984). *Grammar*. Penguin.

[165] Palmer, S. and Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, 1(1):29–55.

[166] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

[167] Pedersen, T. and Bruce, R. (1998). Knowledge lean word-sense disambiguation. In *Proceedings for the 15th National Conference on Artificial Intelligence (AAAI)*, pages 800–805.

[168] Pianta, M. (2013). Democracy lost: The financial crisis in europe and the role of civil society. *Journal of Civil Society*, 9(2):148–161.

[169] Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. *ACM SIGKDD Explorations Newsletter*, 1(2):59–61.

[170] Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2, pages 683–691.

[171] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968.

[172] Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7. Association for Computational Linguistics.

[173] Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 109–117.

[174] Remus, S. and Biemann, C. (2013). Three knowledge-free methods for automatic lexical chain extraction. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 989–999.

[175] Risch, J., Kao, A., Poteet, S., and Wu, Y.-J. (2008). *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, volume 4404 of *LNCS*, chapter Text Visualization for Visual Text Analytics, pages 154–171. Springer.

[176] Rohde, H. and Horton, W. (2010). Why or what next? Eye movements reveal expectations about discourse direction. Talk at the 23rd Annual CUNY Conference on Human Sentence Processing, New York.

[177] Rönnqvist, S., Sarlin, P., Eklund., T., and Back, B. (2012). Mapping bank interrelations in financial discussion. Poster presentation at the Eleventh International Symposium on Intelligent Data Analysis (IDA 2012), Helsinki.

[178] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386.

[179] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[180] Rutherford, A. and Xue, N. (2016). Robust non-explicit neural discourse parser in english and chinese. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL) Shared Task*, pages 55–59.

[181] Rutherford, A. T., Demberg, V., and Xue, N. (2016). Neural network models for implicit discourse relation classification in english and chinese without surface features. *arXiv preprint arXiv:1606.01990*.

[182] Schenk, N. (2018). *Retrieving Implicit Relations from Text. Hidden Semantics and Natural Language Processing*. PhD thesis, Goethe University Frankfurt am Main. Forthcoming.

[183] Schenk, N. and Chiarcos, C. (2017). Resource-lean modeling of coherence in commonsense stories. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 68–73.

[184] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

[185] Schütze, H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796.

[186] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343.

[187] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[188] Smith, A., Lee, T. Y., Poursabzi-Sangdeh, F., Boyd-Graber, J., Elmqvist, N., and Findlater, L. (2017). Evaluating visual representations for topic understanding and their effects on manually generated labels. *Transactions of the Association for Computational Linguistics*, 5:1–16.

[189] Socher, R. and Manning, C. (2013). Deep learning for natural language processing (without magic). Keynote at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013). http://nlp.stanford.edu/courses/NAACL2013/.

[190] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–161.

[191] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[192] Soo, C. (2013). Quantifying animal spirits: news media and sentiment in the housing market. *Ross School of Business Paper No. 1200*.

[193] Sowa, J. F. (1992). *Encyclopedia of Artificial Intelligence*, chapter Semantic Networks. Wiley, 2nd edition.

[194] Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37.

[195] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112.

[196] Tai, K., Socher, R., and Manning, C. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

[197] Tanev, H., Piskorski, J., and Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems (Proc. NLDB 2008)*, volume 5039 of *LNCS*, pages 207–218. Springer.

[198] Tomiyama, T., Umeda, Y., and Kiriyama, T. (1994). A framework for knowledge intensive engineering. In *Computer Aided Systems Theory – CAST'94*, number 1105 in LNCS, pages 123–147. Springer.

[199] Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, 255(5):114–125.

[200] Trevor Hastie, Robert Tibshirani, J. F. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

[201] Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press.

[202] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

[203] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460.

[204] Tutte, W. T. (1963). How to draw a graph. *Proceedings of the London Mathematical Society*, 3(1):743–767.

[205] van Dijk, T. A. (1997). *Discourse as Structure and Process*, chapter The Study of Discourse, pages 1–34. SAGE Publications.

[206] Van Landeghem, S., Hakala, K., Rönnqvist, S., Salakoski, T., Van de Peer, Y., and Ginter, F. (2012). Exploring biomolecular literature with evex: Connecting genes through events, homology, and indirect associations. *Advances in Bioinformatics*, 2012(Article ID 582765):12. Special issue on Literature-Mining Solutions for Life Science Research.

[207] Vickers, D., Lee, M. D., Dry, M., and Hughes, P. (2003). The roles of the convex hull and the number of potential intersections in performance on visually presented traveling salesperson problems. *Memory & Cognition*, 31(7):1094–1104.

[208] Wang, J. and Lan, M. (2016). Two end-to-end shallow discourse parsers for English and Chinese in CoNLL-2016 shared task. In *Proceedings of the Conference on Computational Natural Language Processing (CoNLL) Shared Task*, pages 33–40.

[209] Wang, W. and Hua, Z. (2014). A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[210] Ware, C. (2004). *Information Visualization: Perecption for Design*. Elsevier.

[211] Ware, C. (2008). *Visual Thinking for Design*. Elsevier.

[212] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[213] Webber, B. L. (2004). D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779.

[214] Weiss, G. and Bajec, M. (2016). Discourse sense classification from scratch using focused RNNs. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL) Shared Task*, pages 50–54.

[215] Werbos, P. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University.

[216] Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT Press.

[217] Wilson, T. D. (1984). The cognitive approach to information-seeking behaviour and information use. *Social Science Information Studies*, 4(2-3):197–204.

[218] Wong, P. C. and Thomas, J. (2004). Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21.

[219] Wren, J. D., Bekeredjian, R., Stewart, J. A., Shohet, R. V., and Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20(3):389–398.

[220] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Learning Representations*.

[221] Xue, N., Ng, H. T., Rutherford, A., Webber, B., Wang, C., and Wang, H. (2016). CoNLL 2016 Shared Task on multilingual shallow discourse parsing. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL) Shared Task*, pages 1–19.

[222] Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43.

[223] Zaki, M. J. and Jr., W. M. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

[224] Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

[225] Zhang, B., Su, J., Xiong, D., Lu, Y., Duan, H., and Yao, J. (2015a). Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2230–2235.

[226] Zhang, L., Li, L., and Li, T. (2015b). Patent mining: A survey. *ACM SIGKDD Explorations Newsletter*, 16(2):1–19.

[227] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2.

[228] Zhou, Y. and Xue, N. (2012). PDTB-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 69–77.

[229] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4):479–493.

Paper I

Bank networks from text: Interrelations, centrality and determinants

S. Rönnqvist and P. Sarlin (2015). *Quantiative Finance*, 15(10):1619–1635

Quantitative Finance, 2015
Vol. 15, No. 10, 1619–1635, http://dx.doi.org/10.1080/14697688.2015.1071076

Routledge
Taylor & Francis Group

Bank networks from text: interrelations, centrality and determinants

SAMUEL RÖNNQVIST*† and PETER SARLIN‡§

†Turku Centre for Computer Science – TUCS, Department of Information Technologies, Åbo Akademi University, Turku, Finland
‡RiskLab Finland, Arcada University of Applied Sciences, Helsinki, Finland
§Department of Economics, Hanken School of Economics, Helsinki, Finland

(*Received 15 June 2014; accepted 3 February 2015*)

In the wake of the still ongoing global financial crisis, bank interdependencies have come into focus in trying to assess linkages among banks and systemic risk. To date, such analysis has largely been based on numerical data. By contrast, this study attempts to gain further insight into bank interconnections by tapping into financial discourse. We present a text-to-network process, which has its basis in co-occurrences of bank names and can be analysed quantitatively and visualized. To quantify bank importance, we propose an information centrality measure to rank and assess trends of bank centrality in discussion. For qualitative assessment of bank networks, we put forward a visual, interactive interface for better illustrating network structures. We illustrate the text-based approach on European Large and Complex Banking Groups during the ongoing financial crisis by quantifying bank interrelations and centrality from discussion in 3M news articles, spanning 2007Q1 to 2014Q3.

Keywords: Bank networks; Information centrality; Systemic risk; Text analysis

1. Introduction

The global financial crisis has brought several banks, not to say entire banking sectors, to the verge of collapse. This has not only resulted in losses for investors but also costs for the real economy and welfare at large. Considering the costs of banking crises, the recent focus of research on financial instabilities is well motivated. First, real costs of systemic banking crises have been estimated to average at around 20–25% of GDP (e.g. Dell'Ariccia *et al*. 2008, Laeven and Valencia 2010). Second, data from the European Commission illustrate that government support for stabilizing banks in the European Union (EU) peaked at the end of 2009. The support amounted to 1.5 trl, which is more than 13% of EU GDP. The still ongoing financial crisis has stimulated a particular interest in systemic risk measurement through linkages, interrelations and interdependencies among banks. This paper advances the literature by providing a novel measure of bank linkages from text and bank importance through information centrality.

Most common sources for describing bank interdependencies and networks are based upon numerical data like interbank asset and liability exposures or payment flows, and co-movements in market data (e.g. equity prices, CDS spreads and bond spreads) (see Cerutti *et al*. 2012). While these direct and indirect linkages complement each other, they exhibit a range of limitations. Even though in an ideal world bank networks ought to be assessed through direct, real linkages, interbank data between banks' balance sheets are mostly not publicly disclosed. In many cases, even regulators have access to only partial information, such as lack of data on pan-European bank linkages despite high financial integration. In this vein, a commonly used source of data descends from interbank payment systems (see Soramäki *et al*. 2007) but is again only accessible for a limited set of regulators. It is also worth noting that real exposures, as they are measured for individual markets, are oftentimes highly biased towards the business model of a bank, such as investment or depository functions. Market price data, while being widely available and capturing other contagion channels than those in direct linkages between banks (Acharya *et al*. 2012), assume that asset prices correctly reflect all publicly available information on bank risks, exposures and interconnections. Yet, it has repeatedly been shown that securities markets are not always efficient in reflecting information about stocks (e.g. Malkiel 2003). Further, co-movement-based approaches, such as that by Hautsch *et al*. (2013), require large

*Corresponding author. Email: sronnqvi@abo.fi
This paper is accompanied by supplementary interactive interfaces: http://risklab.fi/demo/textnet/ (for a further discussion of the VisRisk platform, see Sarlin (2014)).

amounts of data, often invoking reliance on historical experience, which may not represent the interrelations of today. Also, market prices are most often contemporaneous, rather than leading indicators, particularly when assessing tail risk. It is not an entirely straightforward task to separate the factors driving market prices in order to observe bilateral interdependence (Borio and Drehmann 2009).

Big data has emerged as a central theme in analytics during the past years. Research questions of big data analytics arise not only from massive volumes of data, or speeds at which data are constantly generated, but also from the widely varying forms, particularly unstructured textual data, that in themselves pose challenges in how to effectively and efficiently extract meaningful information (Dhar 2013). This paper treats the text mining aspect, as it proposes an approach to relationship assessment among banks by analysing how they are mentioned together in financial discourse, such as news, official reports and discussion forums. The idea of analysing relations in text is in itself simple but widely applicable. It has been explored in various areas; for instance, Özgür *et al*. (2008) study co-occurrences of person names in news, and Wren *et al*. (2004) extract biologically relevant relations from research articles. These approaches can be used to construct social or biological networks, using text as the intermediate medium of information. Our contribution lies in proposing this text-based approach to the study of bank interrelations, with emphasis on analysis of the resulting bank network models and ultimately quantifying a bank's importance or centrality.

Our approach may be compared to the above discussed, more established ways of quantifying bank interdependence, such as interbank lending and co-movement in market data. While not measuring direct interdependence, it has the advantage over interbank exposures by relying upon widely available data, and over co-movements in market data by being a more direct measure of an interrelation. By contrast, our approach serves to shed light on banks' relationships in the view of public discussion, or of information overall, depending on the scope of textual data. It may serve as a way of tapping into the wisdom of the crowd, while offering a perspective different from previous methods, especially considering the presence of rich, embedded contextual detail. Rather than an ending point, this sets a starting point from which further study may focus more extensively on the context of occurrences and more sophisticated semantic analysis. This allows to better understand factors driving interrelations, and overall centrality.

In this paper, we assess European Large and Complex Banking Groups (LCBGs) using the text-based approach for quantifying bank interrelations from discussion in the news. A co-occurrence network is derived from 3 million articles, published during 2007Q1 to 2014Q3 in the Reuters online news archive. Beyond only quantifying bank interrelations, we also provide means for quantitative and qualitative assessment of networks. To support quantification of bank importance, we propose an information centrality measure to rank and assess trends of bank centrality in discussion, which relates to the information channel in the analysis of interconnected, and potentially systemic, financial risk. In contrast to common shortest-path-based centrality measures, information centrality captures effects that might propagate aimlessly by accounting for parallel paths. Thus, rather than direct financial exposures, we provide a representation of the channel for potential informational contagion, as well as other common factors leading to co-occurrence in discussion, such as overlapping portfolios and exposure to common exogenous shocks. To support a qualitative assessment of the bank networks, we put forward a visual, interactive interface for better illustrating network structures. This concerns not only an interface to network models, but also an interactive plot to better communicate quantitative network measures.†

The co-occurrence network illustrates relative prominence of individual banks, and segments of more closely related banks. The systemic view acknowledges that the centrality of a bank in the network is a sign of importance, and not necessarily its size (cf. too central to fail by Battiston *et al*. (2012)). The dynamics of the network, both local and global, reflect real-world events over time. The network can also be utilized as an exploratory tool that provides an overview of a large set of data, while the underlying text can be retrieved for more qualitative analysis of relations.

To better understand what drives information centrality, and how it ought to be interpreted, we explore determinants of the centrality measure. We investigate a large number of bank-specific risk drivers, as well as country-specific macro-financial and banking sector variables, as well as control for variables measuring bank size. Further, we also assess the extent to which information centrality explains banks' risk to go bad, and compare it to more standard measures of size. Even though bank size is a key factor explaining information centrality, we show that centrality is not a direct measure of vulnerability. This implies that the centrality measure is not biased by the nature of business activities or models, which potentially impacts bank vulnerability (e.g. asset size or interbank-lending centrality). Rather than a narrow, direct measure of interconnectedness, we are capturing systemic importance of a bank more broadly, in terms of connectivity expressed in financial discourse. Yet, while the rich nature of textual data provides possibilities to more specifically query and define interrelationships and other potentially interesting details on banks, interpreting the semantics by computational methods is often challenging. To this end, we also discuss different ways of analysing text-based networks, laying forward some ideas on future directions in their study.

The following section explains the data and methods we use to construct and analyse bank networks from text, whereas section 3 discusses the results of the experiments on textual data, including both qualitative and quantitative analysis. Before a concluding discussion on text-based networks, section 4 assesses determinants of information centrality.

2. Bank networks from text: data and methods

This section provides a discussion of the text-to-network process, both generally and from the viewpoint of the study in this paper. First, we detail the particular text data and choice of banks to be studied. Having established this, we turn to the

†The interactive interfaces are provided as web-based implementations: http://risklab.fi/demo/textnet/. For a further discussion of the VisRisk platform see Sarlin (2014).

process of text analysis and construction of bank co-occurrence networks. This is followed by discussion on the analysis of such networks, including both quantitative and qualitative analysis.

2.1. Data and target banks

Through digitized economic, social and academic activities, we are having access to ever increasing amounts of textual data. While vast amounts of textual data are readily available, there is nothing that assures increases in precision and quality of data. Analytics of big data is increasingly a search for needles in haystacks, where choices in data source, collection methods as well as pre-processing set-ups all need to be carefully directed in order to pick up desired signals. Likewise, when tapping into financial discourse, one needs to clearly narrow the context of collected data and targeted entities of interest, beyond the choice of data source.

The text data we use in this paper are newly collected from Reuters online news archive. News text presents a rather formal type of discourse, which eases interpretation of extracted relations, as opposed to more free-form, user-generated online discussion as explored in earlier work by Rönnqvist and Sarlin (2014). We focus on major consumer banks within Europe, classified by the European Central Bank (2013) as LCBGs, of which 15 are also classified as Globally Systemically Important Banks (G-SIBs) by the Financial Stability Board (2013). See table A1 in the appendix for a list of LCBGs and G-SIBs and the naming convention used in this paper. The period of study is 2007Q1–2014Q3, for which the news archive contains 6.7M articles. We base our analysis on a 45% random sample of articles comprising of 3.0M articles (1.5B words).

The text analysis is based on detecting mentions of bank names in the articles. We look at a set of 27 banks: 5 British, 5 French, 4 German, 4 Spanish, 3 Dutch, 2 Italian, 2 Swiss, 1 Swedish and 1 Danish bank. In order to mitigate a geographical sampling bias, we use the US edition of the Reuters news archive, as no single European edition is available, but rather national editions for only the largest countries.

The chart in figure 1 provides an overview of the trends in total news article volume, as well as the volume of bank name occurrences. Out of all articles, 5.4% mention any of the targeted banks, on average. The volume is relatively low in the beginning of 2007, i.e. the start of the archive. Mentions of banks reach a peak in early 2008, after which it fluctuates between 60k and 110k articles per quarter.

2.2. From text to bank networks

With plain text as a starting point, and relationship assessment as an objective, we analyse co-mentions in financial discourse. Extracting occurrences and co-occurrences from text is the initial step. The relationships are constituents of co-occurrence networks, whose properties can be assessed through both quantitative and visual analysis. Figure 2 provides an overview of the process of transforming text into network models that lend themselves to analysis.

To construct the network, we scan the text for occurrences of bank names to detect and register mentions of those banks. Scanning is performed using patterns manually designed and tested to match with as high accuracy as possible. Generally, the use of manually designed patterns for information extraction in text tends to have high precision but lower recall, but we expect that the reasonably standardized form of discourse we use should mitigate a loss in recall. The pattern for each bank is specified as a set of regular expressions targeting common naming variants such as full name, abbreviations, synonymous names, names of subsidiaries, historical names and spelling variations. The regular expressions are developed and tested iteratively on data to optimize accuracy, going from broader patterns towards higher precision with retained recall.

Co-occurrence analysis is computationally very efficient and versatile in terms of language, compared to more technically sophisticated relation extraction techniques (e.g. based on dependency parsing Bunescu and Mooney 2005), while it offers worse precision of relations. Using co-occurrence-based relation extraction, lets us process billions of words on standard architecture serially in the order of hours, and we assume the substantial data volume to partially compensate for the noisier relation extraction. The framework is language independent and works equally well on English language news as, for instance, on Finnish online discussion (Rönnqvist and Sarlin 2014).

A co-occurrence relation is formed by two bank names occurring in the same context. In the present case, we define the scope of the context as a 400-character sliding window in the text, whereas a wider scope would require less data but increase noise as any co-occurrence is less likely to represent a meaningful relation. In the process, a context is checked for co-occurrence candidates as follows. A context is scanned for substrings matching the defined regular expressions, and a bank occurrence is registered by associating the matching pattern with its corresponding bank. Multiple occurrences of a single bank are counted only once per context, ignoring presumed meaningless repetitions, but an occurrence may participate in multiple relations. A context containing two or more banks yields one or more pair-wise co-occurrence relations. Thus, derived from the set of matches $M \subset \mathbb{N}$ (indexed by bank) in context c, we define the set of co-occurrence relations R as:

$$R_c = \{r \mid r \in M_c \times M_c \wedge r_1 < r_2\}$$

However, we disqualify contexts with more than five banks, as they are likely to be listings that would result in marginally meaningful relations. These design decisions should be adjusted and tested for each new data source, to obtain less noisy results.

We aggregate co-occurrences over time to form links that are weighted by the absolute co-occurrence count during a period (e.g. a quarter). These links form a dynamic network, a series of cross-sectional networks, which allows the extracted relations to be studied using methods for analysis of complex networks. In the network, banks form nodes (or vertices), and aggregated co-occurrence relations form their links (or edges). To extract meaningful quantitative measures of co-occurrence networks, measures designed for weighted networks need to be used. Nevertheless, most conventional network analysis methods are designed for binary (unweighted) networks only (Opsahl *et al.* 2010), which calls for some form of transformation of the network if these measures are to be used, such as by filtering

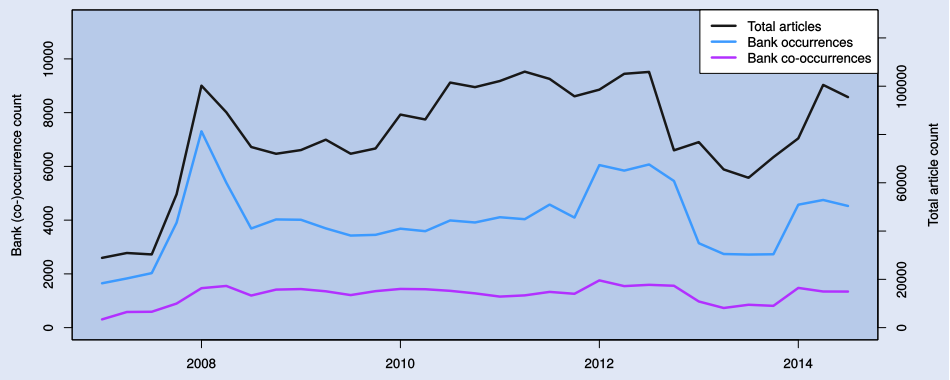


Figure 1. Volumes of all news articles and bank name occurrences over time.

Figure 2. Text-to-network process: (1) Occurrences of bank names are detected in source text, (2) pair-wise co-occurrence relations are extracted between occurrences within a context, and (3) relations aggregated over a time interval form a co-occurrence network. A resulting network can be analysed with (4a) quantitative measures capturing some interesting features, and (4b) qualitative analysis through visual exploration of the network, its neighbourhoods and connectivity of individual nodes.

out very weak connections. While unfiltered networks are more sensitive to noise when using binary measures, low-frequency co-occurrences may be of particular interest, as they are more likely to represent novel information. In order not to lose detail, it is highly motivated to use weighted networks and measures that account for link weights. Larger sample size or longer aggregation intervals increase the co-occurrence count, i.e. the weights of the cross-sectional networks, and will affect many network measures (including information centrality discussed later); as the networks are directly comparable among cross sections this is however not an issue.

Although quantitative analysis of networks provides means to better understand overall properties of networks, they as any aggregate measure most often lack in detail. Hence, network visualization supports not only detailed analysis of network structure and constituents, but also further details as demanded. In the following subsection, we further discuss both quantitative measurement of network properties and visualization as a support in their analysis.

2.3. Network analysis

Network models are commonly rather complex and rich in information. They can be analysed in many different ways to gain insight into the nature of the underlying phenomenon, the bank connectivity landscape in our case. We first discuss analysis of the networks at a global, descriptive level, to describe properties of the co-occurrence networks through common network measures. Later, we concentrate on the concept of centrality and a few ways of quantifying it in our type of network, with the study of systemic risk in mind. Finally, we discuss network visualization as a means for interactive exploration.

2.3.1. Global properties. A commonly cited property of real-world networks is that the average distance between nodes is very small relative to the size of the network, lending them the name ‘small-world’ networks (Watts and Strogatz 1998). Short distances have a functional justification in most types of network, as it increases efficiency of communication, while there also is a general tendency towards short average distances among non-regular networks. These networks have varying *degree*, i.e. the number of links per node, the distribution of which is a typical way of profiling empirical networks. Networks that have evolved through natural, self-organizing processes, such as communications, social, biological and financial networks, tend to exhibit degree distributions that follow a power law. These so-called *scale-free* networks evolve through processes of preferential attachment, where the likelihood of a node receiving a new link is proportional to its current degree (Barabási and Albert 1999).

Jackson and Rogers (2007) distinguish two archetypes of natural networks, described by power-law degree distributions and exponential degree distributions, respectively. They argue

that, in fact, empirical networks generally exhibit hybrid distributions, between power law and exponential, as they are formed through mixed processes of preferential attachment and attachment with uniform probability. The latter process still generates highly heterogeneous exponential distributions, as established nodes have greater chance over time at growing well embedded into the network. By either process, some nodes are bound to be more influential than others, and mapping the levels of influence in the system is our main interest. To profile the co-occurrence networks, the average shortest paths and degree distributions can indicate how small-world and scale-free they are. In the latter case, as we are interested in accounting for the link weighting, we study the distribution of strength, i.e. weighted degree calculated as the sum of weights per node (as Barrat *et al*. 2004 propose).

Other typical ways of characterizing structure focus on network density and modularity. For instance, a clustering coefficient can measure the probability that triplets of connected nodes in binary networks form triangles, providing a measure of density that can be conditioned on degree, etc. Networks may consist of several modules or communities, i.e. subnetworks more densely connected to each other than to other parts. Although such characteristics can be studied by quantitative means, it is not of particular interest for the current news-based bank networks. However, we will briefly consider these qualities based on visual analysis in section 3.2.

2.3.2. Centrality. Following the initial profiling of the whole network, we turn the focus towards the concept of node centrality. A central node holds a generally influential position in a network; a centrally located bank is likely to be systemically important, as it stands to affect a large part of the network directly or indirectly in case of a shock (negative or positive). There is, however, a range of ways to quantify centrality, the most common measures being degree centrality (i.e. fraction of nodes directly linked) and the shortest-path-based closeness centrality and betweenness centrality. We adapt degree centrality to our weighted networks, using strength as a direct measure of centrality. Closeness and betweenness centrality can also incorporate link weight into the calculation of shortest path, by means of the Dijkstra's shortest-path algorithm (Dijkstra 1959) that interprets weights as distances between nodes. Since co-occurrence networks represent tighter connections (i.e. more co-occurrences) by higher weights, it is necessary to invert the weights before calculation, as proposed by Newman (2001).

Borgatti (2005) points out that a common mistake in the study of network centrality is to neglect to consider how flow in the system is best modelled. The common shortest-path-based centrality measures make implicit assumptions that whatever is passing from a node to the surrounding network does so along optimal paths, such as in routing networks of goods and targeted communication. Arguably, a more realistic intuition for influence of a node, in cases where effects might propagate aimlessly, such as any type of contagion, is one that accounts for parallel paths that may exist.

Along these lines, we study a closeness centrality measure that models the flow of information in such a manner, called *information centrality* (Stephenson and Zelen 1989) (also known as current flow closeness centrality Brandes and Fleischer 2005). Information centrality, which seeks to quantify the information that can pass from a node to the network over links whose strength determine level of loss in transmission, is defined as

$$I(i) = \frac{n}{nC_{ii} + \sum_{j=1}^{n} C_{jj} - 2\sum_{j=1}^{n} C_{ij}} \tag{1}$$

where n is the number of nodes and the weighted pseudo-adjacency matrix is defined as

$$C = B^{-1},\ B_{ij} = \begin{cases} 1 + S(i), & \text{if } i = j \\ 1 - w_{ij}, & \text{otherwise} \end{cases}$$

where w is the link weight (0 for unlinked nodes) and $S(i)$ is the strength of node i. This allows us to measure the centrality or influence of bank i in public discourse, which relates to a very general-purpose measure of connectedness in discussion. When relating to systemic risk, we aim at capturing the information channel when analysing interconnected financial risk. Thus, rather than direct financial exposures, we provide a representation of the channel for potential informational contagion, as well as other common factors leading to co-occurrence in discussion, such as overlapping portfolios and exposure to common exogenous shocks.

Centrality as a measure of a node's relative importance is interesting, yet changes in centrality add another dimension. We study networks of quarterly cross sections of the data, in order to calculate and compare centralities over time.

When the data are split by shorter intervals less frequent parts will inevitably become disconnected from the main network component. Information centrality is quite sensitive to the resulting fluctuations in component size, while the more central nodes start to correlate strongly. We propose a method to stabilize the centrality measurement by applying Laplace smoothing to the link weights before calculation of information centrality. The weight of each existing link is increased by a small constant (e.g. 1.0), while links are added between all other nodes and weighted by the same constant. Formally, $w'_{ij} = w_{ij} + \alpha$, where $w_{ij} = 0$ if i and j are not connected. The reasoning is that operating on a limited sample of links, we want to discount some probability for unobserved links (between known nodes), to lessen the influence that the difference between non-occurring (unobserved) links and single-occurrence links has on centrality. This type of additive smoothing has similarly been applied in language modelling (Chen and Goodman 1999) but is generally applicable to smoothing of categorical data.

The choice of the smoothing parameter α is dependent on the study objective: modest levels (e.g. 0.1) retain more information on global changes in centrality, whereas higher levels (e.g. 1.0 or more) accentuate relative differences among nodes. Section 3.1 discusses the effects of different levels of smoothing, based on visual assessment as well as measures of variance. On the one hand, the effect of smoothing can be quantified through the average variance in information centrality per node over time periods (T):

$$V = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|T|}\sum_{t\in T}\left(I'_t(i) - \mu_i\right)^2\right) \tag{2}$$

where μ_i is the mean node centrality over time and I' is smoothed and min-max normalized I over all t and i. This variance should decrease with increased smoothing. On the other

hand, the relative spread of nodes that is expected to increase with higher levels of smoothing can be similarly measured based on variance among nodes in a cross section, rather than among cross sections for a node. The average is then formulated as:

$$V' = \frac{1}{|T|} \sum_{t \in T} \left(\frac{1}{n} \sum_{i=1}^{n} \left(I'_t(i) - \mu_i \right)^2 \right) \quad (3)$$

2.3.3. Visual analysis. While quantitative network analysis plays a vital role in measuring specific aspects of interest in a precise and comparable fashion, network visualization can provide useful overview and exploratory capabilities, communicating general structure as well as local patterns of connectivity. The visual analytics paradigm aims at supporting analytical thinking through interactive visualization, where interaction is the operative term (Keim *et al.* 2008). Through a tight integration between the user and the data model, users are enabled to explore and reason about the data. In the case of our dynamic networks, interaction capabilities for navigating between cross sections and further exploring network structure provide a setting for qualitative analysis of the information-rich models.

Force-directed layouting is often used to apply spatialization of network nodes, that is, to place the nodes in a way that overall approximates node distances to their corresponding link strengths, thereby seeking to uncover the structure of the network in terms of more and less densely connected areas and their relation. Still, force-directed layouts quickly turn uninformative or ambiguous as the networks become too dense, including cases of weighted networks with few strong but many weak connections. Although network visualization with force-directed layouting often does not scale well to analysis of big networks, it still can be a useful tool when used properly. In the case of our bank co-occurrence network, it produces decent visualizations for cross sections of the data-set, while stricter filtering of co-occurrences will produce a more sparse network that is less cluttered. We use the D3 force algorithm (Bostock *et al.* 2011) for layouting.

The dynamics of the network can be studied by visualizing cross-sectional networks in a series, where the positioning is initialized by the previous step and optimized according to the current linkage, as to provide continuity that helps in the visual exploration of network evolution. User interaction plays a vital role not only by allowing to navigate across time, but also by allowing interaction with the positioning algorithm, letting the user acquire a more direct understanding of the structures and details in the data. Force-directed layouting on more densely linked networks generally finds a locally optimal positioning out of a large number of comparably good solutions. Interaction that lets the user drag nodes to reposition them and a force-directed algorithm that helps to counter-optimize the positioning immediately afterwards gives rise to a collaborative, exploratory way of working with and understanding the data.

Nevertheless, the best setting for visual analysis might be one that combines with quantitative analysis, encoding them visually. For instance, centrality measures can be encoded by node size to enhance the communication of structure provided by network visualization, which can use force-directed layouting or other more regularly structured layouts.

3. Centrality: quantitative and qualitative analysis

This section describes the co-occurrence networks from both a viewpoint of quantitative measures and exploratory visualization. Starting with network measures, we describe network properties in general and information centrality in particular. Then, we turn to visual analysis of the networks and their constituents.

3.1. Quantitative analysis

The volume of bank occurrences is rather stable, apart from a peak centred around 2008Q1 and some fluctuation from 2012 onward. In 2008, the peak in occurrence volume coincides with a peak in total article volume, unlike later during the studied time span when occurrence seems less affected by fluctuating article volume. Interestingly, the 2008 surge in occurrences barely translates into a rise in co-occurrences (or strength), i.e. even though banks are more discussed at the time prior to the outbreak of the crisis, they are not discussed considerably more in close connection to each other. Overall, total article volume and bank occurrences have a Pearson correlation of 0.745. Occurrences and co-occurrences have a correlation of 0.835, which indicates that there is a notable component to co-occurrence volume which is not simply explained by occurrence volume.

From these aggregated counts, we continue by studying the data as a network. As discussed in section 2.3, empirical networks are typically profiled through measures describing certain global properties. The average distance, in terms of number of links, between nodes in the co-occurrence networks is certainly small, at 1.1–1.3, and would justify calling them 'small-world' networks. However, with weighted links, a measure of average distance becomes hardly interpretable. While it is clear that our networks are very tightly connected, the strength distribution depicts the relative differences in node connectivity. Many empirical networks exhibit power-law distributed degree or strength, as a sign of evolution through preferential attachment. Figure 3 shows the cumulative strength distribution of the aggregated network for the entire period, as well as a closely fitted exponential function that hints that our network is exceedingly a product of evolution through uniform attachment. Still, we are able to partially fit power-law functions to the distribution, as the figure highlights with straight lines, which could indicate a hybrid model with a weak preferential attachment component as well. The strength distribution illustrates the high heterogeneity of connections in the network, i.e. some banks are much more associated in discussion than others. However, in order to gain a deeper understanding of a bank's importance to the wider network, we need to look beyond immediate connections as measured by degree/strength distribution or degree/strength centrality (proportional to co-occurrence volume), namely we need to look at information centrality.

We study information centrality for each node over time, using different levels of Laplace smoothing (ranging from 0.0 to 5.0). Figure 4 plots the information centrality values, with a number of example banks highlighted in colour and representative α values. Comparing information centrality with and

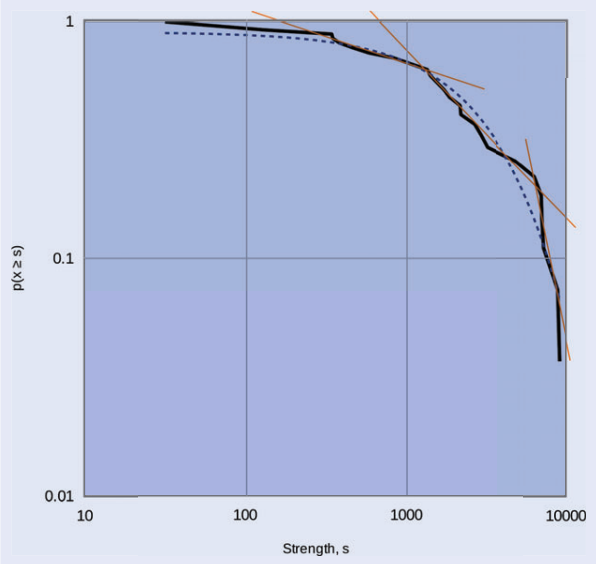


Figure 3. Cumulative strength distribution (weighted degree) of bank co-occurrence network during 2007Q1–2014Q3, showing probability p over node strength x against the upper bound on strength s. Dashed line is a fitted exponential function. Solid straight lines indicate locally fitting power-law functions.

without smoothing visually, we see that different peaks are pronounced: some minor peaks (e.g. during 2009Q2–2011Q4) subside, while others (e.g. prior to 2008Q3 and crisis breakout) are substantially amplified even at low levels of smoothing. Based on its rationale, we interpret that smoothing helps highlight meaningful patterns in information centrality dynamics and generally stabilize the series, while reducing artefacts of changing network size. At higher levels (e.g. $\alpha = 1.0$), peaks become relatively weaker as the distribution of banks evens out on the information centrality scale, so that fewer banks flock at the very top. We aim to measures these respective qualities as V and V' in equations (2) and (3).

The average variance over time V is stationary for very low values of α, with the expected decrease starting at $\alpha = 0.2$ (11% drop from unsmoothed $V = 0.027$) and continuing monotonously with stronger smoothing, directly reflecting its stabilizing nature. Meanwhile, V' signals an increased spread among banks already at $\alpha = 0.01$ (21% over unsmoothed $V' = 0.025$), which reaches a maximum at $\alpha = 1.0$ (94% increase). We conclude that α levels at or slightly above 0.2 are suitable to achieve moderate smoothing that communicates global changes of centrality in this network, whereas 1.0 appears to be the optimal choice when focusing on relative differences in centrality among banks. The regressions in section 4 use information centrality with smoothing at $\alpha = 1.0$, since relative differences in centrality are of particular interest. In our experiments, we also note that V closely follows measures of average covariance of banks over time, supporting the observation that stronger smoothing reduces the originally very strong correlation among the most central banks.

Finally, to test smoothing in relation to sample size, we compare the variance measures when applied to the above discussed 45% sample set to a 20% sample. As Laplace smoothing is a method to mitigate effects of limited sample sizes, we expect relatively stronger effects when applied to a smaller sample. Indeed, $\alpha = 0.1$ results in a 47% drop in V (from 0.03 at $\alpha = 0.0$) at 20% sampling, while higher alpha only has marginal decreasing impact. Even a small α has a strong stabilizing effect on the smaller sample, which in this case contains 1.3M articles. This underlines the fact that working with text data, typically involving very long-tailed distributions, often benefits from big data in terms of size to achieve reliable results, and that smoothing methods are practical for that very reason. In addition, we tested the robustness of information centrality over 10 random samplings (at 30%, $\alpha = \{0.0, 0.1, 1.0\}$) that resulted in standard deviations (relative to the mean centrality of that smoothing level) of 21.2, 8.5 and 3.7% respectively, which further highlights the stabilizing effect of smoothing on sparse data.

The trends of individual banks generally follow the movements of the cross section closely, as increased connectivity in parts of the network strongly affects the rest, since the co-occurrence network is generally very tightly connected. Individual centrality relative to the cross section is generally quite stable. Nevertheless, some changes can be observed that might reflect real-world events. For instance, ABN AMRO has relatively high information centrality in 2007 that decreases afterwards. Royal Bank of Scotland is the most central bank in 2007–2008, whereas it later on is overtaken by, e.g. Barclays and Deutsche Bank. To illustrate the information centrality

Figure 4. Information centrality for banks over time. The charts show different levels of smoothing: none ($\alpha = 0.0$), little ($\alpha = 0.2$) and moderate ($\alpha = 1.0$). A few example banks are highlighted (bank labels are described in table A1 in the appendix).

ranking between banks in more detail, figure 5 shows all values as of 2014Q3.

The smoothed information centrality plots exhibit peaks in both 2008Q1 and 2008Q4, as well as during 2012 and 2014Q1. In 2008Q1, for instance, the peak coincides with the peak in bank occurrence. The fact that co-occurrence stays relatively flat during the same time indicates that the change in information centrality is not so much due to generally strengthened connections, but largely due to change in topology. The peak in the fourth quarter likewise hints at topological shifts following the crisis outbreak, but in this period even bank occurrence is normal. Centralities rise towards 2012, but have subsided substantially in 2013, then coinciding with a similar sharp decrease in bank occurrence. Overall, the correlation between co-occurrence volume and raw information centrality averaged over all nodes is 0.651, hinting at a considerable component other than general co-occurrence volume that we argue is topological, i.e. involving changes of weight distribution over links as well as changes in link structure.

3.2. Visual analysis

As a complement to the discussion on quantitative analysis of the co-occurrence networks, we briefly consider the role of visual network analysis. Our information centrality measurements highlight an interesting pattern in 2008Q2–2008Q4 that we inspect further visually. The second and fourth quarters have relatively high global information centrality, whereas there is a temporary dip in the third quarter. The networks in figure 6 show visualized snapshots of each quarter, where the changes in patterns of connectivity can be studied in more detail. It shows a sparser topology for Q3 than in both Q2 and Q4, as reflected by the measure. In addition, the visualization allows for studying local patterns, e.g. how the connection between the two Scandinavian banks Nordea and Danske Bank (right-side edge of networks) changes.

In general, the networks have a core consisting of the most central banks that does not change drastically over time. The periphery experiences more topological change, but its banks stay mostly in their outside positions. Nevertheless, it is hardly

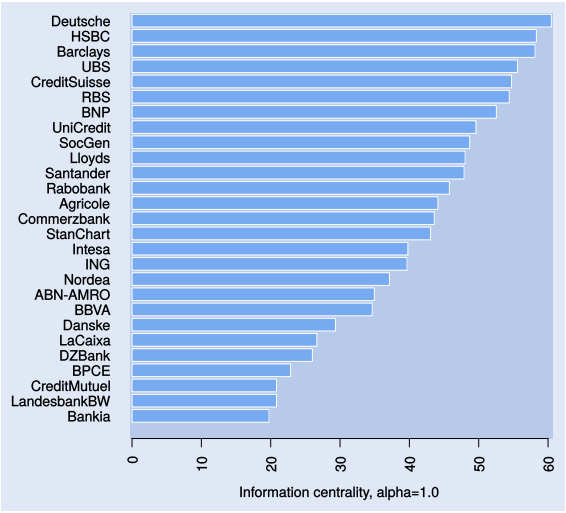


Figure 5. Information centrality ranking for all banks in 2014Q3 (bank labels are described in table A1 in the appendix).

possible to define a strict border between core and periphery, neither by visual inspection nor quantitatively (e.g. by degree or information centrality), rather the nodes appear on a continuum of centrality (cf. figure 5). We may interpret from the force-directed visualization that the network consists of one major module, with the only exception of occasionally disconnected components or single nodes (e.g. La Caixa and Bankia in figure 6(b)). The network is overall very densely connected in terms of binary links.

Even though visual inspection can provide valuable insight, in many cases, it may be hard to reliably and precisely compare changes in specific aspects, such as centrality of single nodes or centralization of the whole network, based on the network visualization. This underlines the importance of backing visual analysis with quantitative measures, for instance, by encoding node size with information centrality or presenting plots of measures in parallel, coordinated views. The combination of both approaches is posed to provide the best possibilities for understanding the properties of the network, through a mixed process of exploration and focused inspection. The visual representations in figure 6 represent information centrality as node size, which in combination with the force-directed node positioning provides support for visually assessing node centrality in more general terms.

4. Determinants of information centrality

Analysis thus far attempted to convince that information centrality captures the notion of system-wide importance of a bank in terms of financial discourse. Yet, little was done to provide a deeper interpretation of what information centrality signifies. This section explores potential determinants of information centrality. We explain centrality with a large number of bank-specific risk drivers, as well as country-specific macro-financial and banking sector variables, beyond controls for bank size. Further, we also assess the extent to which information centrality explains banks' risk to go bad, and compare it to more standard measures of size.

4.1. Data

We complement the textual data, and there from derived centrality measures, with bank-level data from financial statements and banking sector and macro-financial indicators at the country level. This gives us a data-set of 24 risk indicators, spanning 2000Q1 to 2014Q1 for 27 banks, as well as distress events. The definitions of distressed banks follow Betz *et al*. (2014) and are defined based upon the following three categories of events:

- Direct bank failures include bankruptcies, liquidations and defaults.
- Government aid events comprise the use of state support on the asset side, such as capital injections or participation in asset relief programmes (i.e. asset protection or asset guarantees).
- Forced mergers capture private sector solutions to bank distress by conditioning mergers with negative coverage ratios or a parent receiving state aid after a merger.

To measure risk drivers, we make use of CAMELS variables (where the letters refer to Capital adequacy, Asset quality, Management quality, Earnings, Liquidity and Sensitivity to market risk). The Uniform Financial Rating System, informally known as the CAMEL ratings system, was introduced by the US regulators in 1979. Since 1996, the rating system was complemented with Sensitivity to Market Risk, to be called CAMELS. The literature on individual bank failures draws heavily on the

Table 1. Regression estimates on determinants of information centrality.

		Estimates	(1) *Size*	(2) *C*	(3) *A*	(4) *M*	(5) *E*	(6) *L*	(7) *S*	(8) *CAMELS*	(9) *Nosize*	(10) *Bank +country*
Bank-specific indicators		Intercept	9.06***	9.06***	9.23***	9.05***	9.06***	9.43***	9.03***	9.31***	9.32***	9.20***
	I[a]	Total assets	1.36***	1.80***	1.57***	1.50***	1.28***	2.64***	1.41***	3.53***		3.29***
		Total deposits	1.37***	0.93*	0.93**	1.20***	1.43***	−0.55	1.26***	−1.66**		−1.53*
	C[a]	Total equity to total assets		0.50*						0.11	−0.28	0.39
	A[a]	Reserves for loan losses to impaired assets			−0.52**					−1.79**	−1.84**	−0.92
		Loan loss provisions to total loans			0.86***					0.69**	1.06***	0.42
	M[a]	Operating costs to operating income				−0.43*				−0.52	−0.37	−1.20**
	E[a]	Return on assets					−0.25			−0.03	0.37	0.08
		Return on equity					−0.05			−0.41*	−0.55*	−0.08
	L[a]	Interest expenses to total liabilities						−0.96***		−0.82***	−1.49***	−1.12***
		Deposits to funding						1.12***		1.58***	0.55*	1.30***
		Net-short term borrowing						−0.38*		−0.22	−0.10	0.44.
	S[a]	Share of trading income							0.27	−0.32	−0.03	−0.64.
Country-specific banking sector indicators		Total assets to GDP										1.62***
		Non-core liability growth										0.41*
		Debt to equity										−0.44
		Debt securities to liabilities										0.81
		Mortgages to loans										0.52
		Loans to deposits										−2.46***
Country-specific macro-financial indicators		Real GDP growth										−0.26
		Inflation										−0.82**
		Stock price growth										−1.28***
		House price growth										0.18
		Long-term government bond yield										−1.05***
		International investment position to GDP										0.75
		Government debt to GDP										0.23
		Private sector credit flow to GDP										0.23
		Multiple R^2	0.24	0.25	0.27	0.25	0.25	0.30	0.24	0.33	0.19	0.46
		Adjusted R^2	0.24	0.25	0.27	0.25	0.24	0.29	0.24	0.31	0.17	0.43

Notes: For standardized coefficients, the explanatory variables have been transformed to have zero mean and unit variance.
Signif. codes: '***', 0.001; '**', 0.01; '*', 0.05; '.', 0.10.
[a]The letters of CAMELS refer to Capital adequacy, Asset quality, Management, Earnings, Liquidity and Sensitivity to market risk, and I refers to bank size.

Table 2. Early-warning models with information centrality.

		Estimates	(1) IC	(2) IC + assets	(3) IC + deposits	(4) IC + size	(5) IC + CAMELS	(6) Assets + CAMELS	(7) Deposits + CAMELS	(8) Size + CAMELS	(9) IC + all	(10) Assets + all	(11) Deposits + all	(12) All
Bank-specific indicators		Intercept	0.12***	–2.27***	–2.76***	–2.79***	–4.37***	–4.78***	–5.03***	–5.73***	–18.90***	–17.36***	–21.14***	–25.62**
	I[a]	Information centrality	–0.005**	–0.04	–0.004	–0.01	–0.02			0.02	0.07			0.19
		Total assets		0.44*		0.82*		0.61*		2.03*		0.03		0.33
		Total deposits			–0.91***	–1.71***			–0.94**	–3.47**			–2.34*	–2.58
	C[a]	Total equity to total assets					–1.92***	–1.97***	–1.94***	–1.89***	–2.59*	–2.59*	–2.93*	–3.20*
	A[a]	Reserves for loan losses to impaired assets					-1.03	-1.06	–1.54	-2.64*	–0.25	–0.46	0.47	1.57
		Loan loss provisions to total loans					–1.10*	–1.04*	–1.10*	–1.38**	–1.60	–1.52	–3.20*	–3.49.
	M[a]	Operating costs to operating income					–1.36**	–1.39**	–1.36**	–1.31**	–0.86	–1.08	–2.19	-2.00
	E[a]	Return on assets					0.02	–0.04	0.00	0.16	–1.47	–1.38	–3.66*	–4.27*
		Return on equity					–1.17.	–1.02	–1.08.	–1.40*	–1.80	–1.73	–0.25	–0.42
	L[a]	Interest expenses to total liabilities					–0.18	–0.21	–0.16	0.01	–3.10*	–2.76*	–2.82.	–3.70
		Deposits to funding					0.75*	0.78**	1.15***	2.12***	4.81*	4.48*	4.54*	5.28.
		Net-short term borrowing					1.22***	1.44***	1.51***	1.55***	3.78**	3.45**	2.82*	3.32*
	S[a]	Share of trading income					–0.94**	–0.96**	–0.93**	–0.86**	–1.63	-1.66	-2.53.	–2.71.

(*Continued*)

Table 2. (*Continued*)

	Estimates	(1) IC	(2) IC + assets	(3) IC + deposits	(4) IC + size	(5) IC + CAMELS	(6) Assets + CAMELS	(7) Deposits + CAMELS	(8) Size + CAMELS	(9) IC + all	(10) Assets + all	(11) Deposits + all	(12) All
Country-specific banking sector indicators	Total assets to GDP									–8.30*	–7.48*	–8.75*	–10.62*
	Non-core liability growth									1.81	2.09.	2.80.	2.32
	Debt to equity									2.96	2.60	3.60	4.83
	Debt securities to liabilities									-4.22	–3.60	–7.12	–9.46.
	Mortgages to loans									0.93	0.22	0.80	2.59
	Loans to deposits									5.80*	5.25*	7.12*	9.09*
Country-specific macro-financial indicators	Real GDP growth									1.17	1.14	1.22	1.32
	Inflation									0.66	0.58	0.08	0.11
	Stock price growth									–3.98***	–4.06***	–4.88***	–4.92**
	House price growth									–1.76	–1.54	–2.33	–2.53
	Long-term government bond yield									3.81*	3.33**	4.13**	5.64*
	International investment position to GDP									2.77	2.32	3.39	4.47
	Government debt to GDP									–7.81**	–7.35**	–9.49**	–11.11**
	Private sector credit flow to GDP									–4.07**	–3.85**	–4.85**	–5.84**
	Predictive performance												
	AUC	0.63	0.64	0.70	0.75	0.91	0.92	0.92	0.92	0.99	0.99	0.99	1.00
	Usefulness for a policymaker	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$	λ $U_r(\mu)$
Policymaker's preferences[b]	$\mu = 0.6$	1.00 0%	1.00 0%	1.00 0%	0.98 1%	0.97 25%	0.91 39%	0.91 39%	0.91 39%	0.93 82%	0.93 85%	0.93 85%	0.93 85%
	$\mu = 0.7$	1.00 0%	1.00 0%	1.00 0%	0.98 5%	0.91 34%	0.91 53%	0.91 53%	0.91 53%	0.91 85%	0.93 86%	0.93 86%	0.93 86%
	$\mu = 0.8$	1.00 0%	1.00 0%	0.95 7%	0.96 11%	0.91 47%	0.91 63%	0.91 63%	0.91 63%	0.91 87%	0.91 88%	0.92 88%	0.92 88%
	$\mu = 0.9$	0.69 15%	0.77 15%	0.74 30%	0.71 41%	0.87 61%	0.91 71%	0.91 71%	0.91 71%	0.89 91%	0.91 92%	0.92 92%	0.92 93%

Notes: For standardized coefficients, the explanatory variables have been transformed to have zero mean and unit variance. Bolded Usefulness results refer to benchmark preferences of $\mu = 0.9$.
Signif. codes: '***', 0.001; '**', 0.01; '*', 0.05; '.', 0.10.
[a] The letters of CAMELS refer to Capital adequacy, Asset quality, Management, Earnings, Liquidity and Sensitivity to market risk, and I refers to bank size (i.e. importance).
[b] The Usefulness for a policymaker is computed for a threshold λ? that optimizes relative usefulness $U_r(\mu)$ as described in the Appendix.

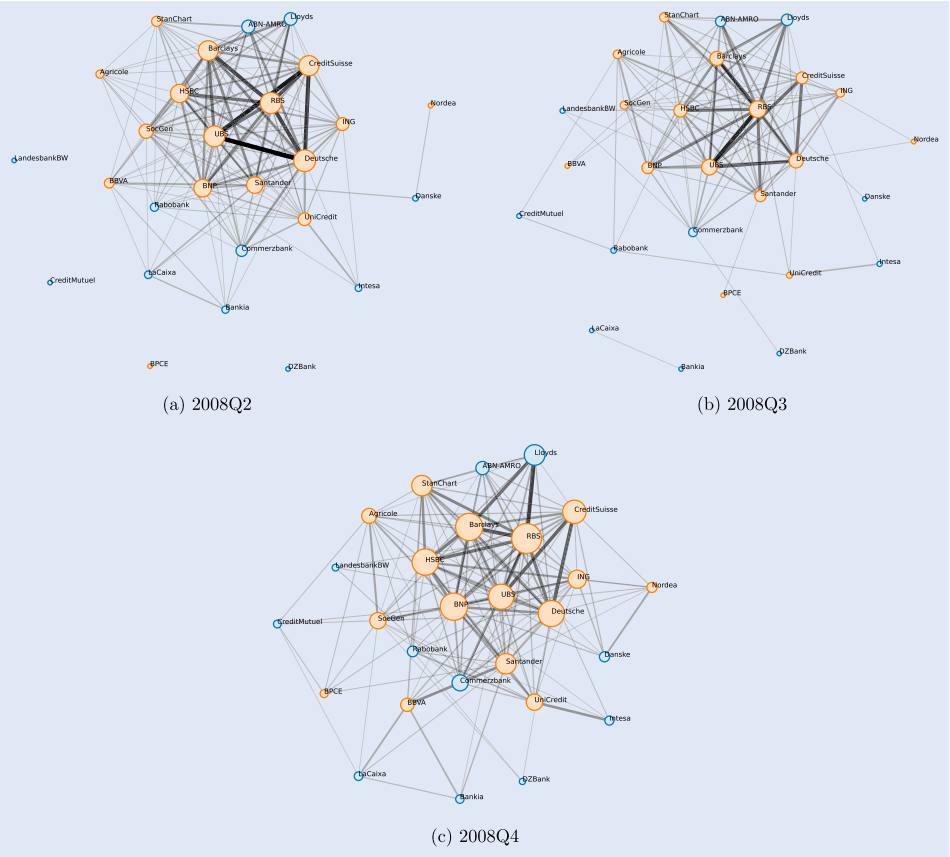


Figure 6. Network visualization for 2008Q2–2008Q4, each showing current topology and link strengths (encoded as opacity and logarithmically scaled line width). Node size is relative to information centrality ($\alpha = 0.2$) and orange colour denotes globally systemically important banks (bank labels are described in table A1 in the appendix).

risk drivers put forward by the CAMELS framework. Further, we complement bank-level data with country-level indicators of risk. One set of variables describes the banking sector as an aggregate, whereas another explains macro-financial vulnerabilities in European countries, such as indicators from the scorecard of the Macroeconomic Imbalance Procedure. All bank-specific data are retrieved from Bloomberg, whereas country-level data come mainly from Eurostat and ECB MFI Statistics.

4.2. What explains information centrality?

The essential question we ask herein is whether more central banks perform or behave differently. Following Bertay *et al.* (2013), who assess whether and to what extent performance, strategy and market discipline depend on standard bank size measures, we conduct experiments in order to better understand what signifies information centrality. In contrast to their study, we control for more standard measures of bank size, in order to capture particular effects of information centrality. Using the above described data, we make use of standard, linear least squares regression models to conduct the following experiments (cf. table 1):

(1) Explain information centrality (IC) with bank size variables (Model 1).
(2) Explain IC with CAMELS variable groups one-by-one, controlling for bank size (Models 2–7).
(3) Explain IC with all CAMELS variables, controlling for bank size (Model 8).
(4) Explain IC with CAMELS and country-specific variables, controlling for bank size (Model 9).

Our experiments show a number of patterns about drivers of information centrality. Table 1 summarizes all regression estimates. First, we show that size measures of total assets

and total deposits statistically significantly explain information centrality. This holds both when included individually and together in regressions. At a 0.1% level, we can show that these size variables relate to centrality, which is in accordance with the nature and aim of the measure.

Second, we also add variable groups from the CAMELS framework to assess which risk factors explain information centrality. When testing groups one-by-one, we find that equity to assets, cost-to-income ratio and net-short-term borrowing are statistically significant at the 5% level, and loan loss provisions to total loans, reserves to impaired assets, interest expenses to liabilities and deposits to funding are statistically significant at the 1% level. Large cost-to-income ratios are expected to reduce individual bank risk, whereas loan loss provisions are expected to increase risk. Yet, the estimates of the liquidity variables—interest expenses to total liabilities and deposits to funding—indicate less risk, as more deposits is expected to be negatively and more interest expenses positively related to bank risk. The relationships of loan loss reserves and share of trading income are potentially ambiguous, as higher reserves should correspond to a higher cover for expected losses but could also proxy for higher expected losses and trading income might be related to a riskier business model as a volatile source of earnings, but investment securities are also liquid, allowing to minimize potential fire sale losses.

Third, when including all size and CAMELS variables, we still find the same variables to be statistically significant, except for all variables significant at the 5% level (i.e. equity to assets, cost-to-income ratio and net-short-term borrowing). When assessing the size variables, assets is consistently a significant predictor, whereas deposits turns insignificant in Model 6 when also including deposits to funding, which is likely to be a result of multicollinearity. Further, the effects of individual risk indicators are unchanged when excluding all bank size variables, except for slight changes in significance levels. Fourth, we complement the bank-specific model with country-level data by also explaining centrality with banking sector and macro-financial variables. Even though this leads to an improvement of R^2 by one-third, this leaves most of the effects unchanged. Notably, liquidity indicators and the cost-to-income ratio remain statistically significant. Out of the country-specific variables, statistically significant predictors are assets to GDP, non-core liability growth, loans to deposits, inflation, stock price growth and sovereign bond yields.

4.3. Information centrality as a risk driver

In the above experiments, we showed that information centrality is partly driven by CAMELS variables, which generally represent different dimensions of individual bank risk. This does not, however, necessarily imply that information centrality is a measure of vulnerability. The next question is whether and to what extent information centrality signals vulnerable banks, particularly when controlling for CAMELS variables.

As we have distress events for the banks, and the above used risk indicators, we can easily test the extent to which information centrality aids in identifying vulnerable banks. By focusing on vulnerable rather than distressed banks, we are interested in periods that precede distress events (e.g. 24 months). In this case, we make use of standard logistic regression to attain a predicted probability for each bank to be vulnerable. This probability is turned into a binary point forecast by specifying a threshold above which we signal vulnerability. This threshold is chosen to minimize a policy-maker's loss function, who has relative preferences between false alarms and missed crises. Also, we provide a so-called usefulness measure that captures the performance of the model in comparison to not having a model (i.e. best guess of a policy-maker). We assume in the benchmark case, the policy-maker to be more concerned about missing a crisis than giving a false alarm, which is particularly feasible for internal signals. See appendix 2 for more details of the evaluation measures.

To test to what extent information centrality signals vulnerabilities, and how it relates to bank size variables, we regress pre-distress events. Hence, as in a standard early-warning setting for banks, we explain periods 24 months prior to distress with logistic regression. Starting out from bank importance variables, we can see in table 2 (Models 1–4) that while none of the variables yield highly valuable predictions, assets and deposits provide more usefulness than information centrality, particularly deposits. The same holds also for statistical significance. Even though the bank size variables were above shown to explain information centrality, we can observe a difference in their relation to risk. Large banks in terms of assets are found to be more vulnerable to distress, whereas large banks in terms of deposits are found to be less so. This is likely to proxy for the business model or activities of a bank, which might be less risky when the focus is on depository functions. Moreover, deposits can be seen as a more stable funding source than interbank market or securities funding. This points to information centrality being a more general measure of interconnectedness, rather than one defined by the underlying focus of the business model. Further, when we add all CAMELS variables to the three importance measures (Models 5–8), both usefulness and statistical significance points to better explanatory power of assets and deposits. Comparing to models with only bank importance variables, this moves usefulness from $U_r(\mu = 0.9) = 41\%$ at its maximum to 61% for information centrality and 71% for assets and deposits. Likewise, when adding all country-specific variables (Models 9–12), we can still observe that the explanatory power of assets and deposits is higher than that for information centrality. At this stage, we have early-warning models that capture most of the available usefulness, by showing a $U_r(\mu = 0.9) \geq 91\%$. When assessing performance with the area under the receiver operating characteristic curve (AUC) (see appendix 2 for a description), we can see that the same conclusions with respect to performance hold.

The implication of the two conducted experiments jointly is that information centrality is highly correlated with bank size, both when measured in total assets and deposits, but not a measure of vulnerability. This indicates that the measure is not biased by business activities or models, which might be a factor impacting the vulnerability of a bank. Rather, we are capturing more broadly importance of a bank in terms of information connectivity in financial discourse. This property, while due to its broad nature may be a disadvantage, provides ample means for measuring interconnectedness and centrality from a wider perspective. It is worth remembering that these

text-based networks are not an ending point but allow further exploration of the semantics of observed connections.

5. Conclusions

The ongoing global financial crisis has brought interdependencies among banks into focus in trying to assess interconnected and systemic risk. This paper has demonstrated the use of computational analysis of financial discussion, as a source for information on bank interrelations. Conventional approaches make use of direct linkages to the extent available and market-based measures as an indirect estimate of interdependence, which both have their limitations, such as non-publicly disclosed information, strong ties to specific business models and deficiencies in the forward-lookingness of co-movements in markets. The approach we put forward may serve as a complement to more established ways of quantifying connectedness and dependence among banks. We have presented a text-to-network process, which has its basis in co-occurrences of bank names and can be analysed quantitatively and visualized. To support quantification of bank importance, we proposed an information centrality measure to rank and assess trends of bank centrality in discussion. Rather than a common shortest-path-based centrality measure, information centrality captures effects that might propagate aimlessly by accounting for parallel paths. Moreover, we proposed a method to stabilize the centrality measure by applying Laplace smoothing to link weights before calculating information centrality. To support a qualitative assessment of the bank networks, we put forward a visual, interactive interface for better illustrating network structures. This concerned not only an interface to network models, but also an interactive plot to better communicate quantitative network measures. Our text-based approach was illustrated on European LCBGs during the ongoing financial crisis by quantifying bank interrelations from discussion in 3.0M news articles, spanning the years 2007–2014 (Q3).

To better understand the interpretation of, and what drives, information centrality, we have explored determinants of the centrality measure. We investigated bank-specific and country-specific risk drivers, as well as control for variables measuring bank size, and also assess the extent to which bank risk is explained by information centrality in relation to more standard measures of size. We have shown that centrality is not a direct measure of vulnerability, despite the fact that it is closely linked to size variables. The conclusions to be drawn from this are that the centrality measure is not biased by the nature of business activities or models, which may impact the vulnerability of banks (e.g. asset size or interbank-lending centrality). Instead, the measure of information centrality is described to capture the importance of a bank in a wider perspective, in terms of information connectivity in financial discourse.

Considering the limitations of the current network and that the underlying data occasionally lead to somewhat hazy patterns difficult to interpret and draw clear conclusions from, we suggest a number of ways these issues could be addressed in future research. One advantage of using text data is the potentially rich semantic information it holds, which can be used to better explain or narrow the relations extracted, thereby facilitating interpretation of the network and the measures applied on top. A disadvantage of applying such filtering might be that it vastly increases the data size requirements, quickly reducing a big but sparse data-set into a rather scarce one. Similarly, in this paper, we have illustrated how, using a data-set of a few million articles, accuracy can still be improved by even more data.

Although textual data provide the basis for studying interrelationships and other potentially interesting details on banks more specifically, its interpretation by computational methods is often challenging. In order to apply filtering by theme to co-occurrence links between banks, we recommend more sophisticated semantic analysis to increase recall. For instance, distributional semantic methods (Turney and Pantel 2010, Mikolov *et al*. 2013) could be used to extend a set of seed keywords, or probabilistic topic modelling (Blei 2012) could be applied to the corpus to identify topics of interest and the related subset of articles. Furthermore, combining sentiment analysis with our bank relation extraction could constitute another interesting way to distinguish the nature of mapped relations. Sentiment analysis has been applied to classify company-related information from financial news in regard to the effect on their stock price (e.g. Malo *et al*. 2013), an approach that could hold considerable potential in the area of systemic risk analysis as well.

Acknowledgements

The authors want to thank Tuomas Peltonen and three anonymous reviewers for insightful comments and discussions. The paper has also gained from presentations of previous versions of it at the following conferences: the 11th International Symposium on Intelligent Data Analysis (IDA'12) on 25–27 October 2012 in Helsinki, Finland, the 19th IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr'14) on 27–28 March 2014 in London, UK, and Arcada Seminar on Current Topics in Business, IT and Analytics (BITA'14) on 13 October 2014 in Helsinki, Finland. The first author gratefully acknowledges the Graduate School at Åbo Akademi University and the second author the GRI in Financial Services, Louis Bachelier Institute, and the Osk. Huttunen Foundation for financial support.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Acharya, V., Pedersen, L., Philippon, T. and Richardson, M., Measuring systemic risk. Discussion Papers, No. 8824, Centre for Economic Policy Research, 2012.

Barabási, A.-L. and Albert, R., Emergence of scaling in random networks. *Science*, 1999, **286**(5439), 509–512.

Barrat, A., Barthelemy, M., Pastor-Satorras, R. and Vespignani, A., The architecture of complex weighted networks. *Proc. Nat. Acad. Sci. U.S.A.*, 2004, **101**(11), 3747–3752.

Battiston, S., Puliga, M., Kaushik, R., Tasca, P. and Caldarelli, G., Debtrank: Too central to fail? Financial networks, the fed and systemic risk. *Sci. Rep.*, 2012, **2**, 541. Available online at: http://www.nature.com/srep/2012/120802/srep00541/full/srep00541.html?WT.mc_id=TWT_SciReports.

Bertay, A.C., Demirgüç-Kunt, A. and Huizinga, H., Do we need big banks? Evidence on performance, strategy and market discipline. *J. Financ. Intermed.*, 2013, **22**(4), 532–558.

Betz, F., Oprica, S., Peltonen, T. and Sarlin, P., Predicting distress in European banks. *J. Bank. Finance*, 2014, **45**, 225–241.

Blei, D.M., Probabilistic topic models. *Commun. ACM*, 2012, **55**(4), 77–84.

Borgatti, S.P., Centrality and network flow. *Soc. Networks*, 2005, **27**(1), 55–71.

Borio, C.E.V. and Drehmann, M., Towards an operational framework for financial stability: "fuzzy" measurement and its consequences. Number 284, Bank for International Settlements, Monetary and Economic Department, 2009.

Bostock, M., Ogievetsky, V. and Heer, J., D3: Data-driven documents. *IEEE Trans. Visual. Comp. Graphics (Proc. InfoVis)*, 2011, **17**(12), 2301–2309.

Brandes, U. and Fleischer, D., Centrality measures based on current flow. *STACS 2005*, 2005, **3404**, 533–544. Available online at: http://link.springer.com/chapter/10.1007/978-3-540-31856-9_44.

Bunescu, R.C. and Mooney, R.J., A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724–731, 2005 (Association for Computational Linguistics: Vancouver).

Cerutti, E., Claessens, S. and McGuire, P., Systemic risk in global banking: What can available data tell us and what more data are needed? Working Papers No. 376, Bank for International Settlements, 2012. Available online at: http://www.bis.org/publ/work376.htm.

Chen, S.F. and Goodman, J., An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.*, 1999, **13**(4), 359–393.

Dell'Ariccia, G., Detragiache, E. and Rajan, R., The real effect of banking crises. *J. Financ. Intermed.*, 2008, **17**(1), 89–112.

Dhar, V., Data science and prediction. *Commun. ACM*, 2013, **56**(12), 64–73.

Dijkstra, E.W., A note on two problems in connexion with graphs. *Numer. Math.*, 1959, **1**(1), 269–271.

European Central Bank, *Financial Stability Review*, November, 2013. Available online at: http://www.ecb.europa.eu/pub/pdf/other/financialstabilityreview201311en.pdf?6fd6719bc2d089c67c09358ebea3be31.

Financial Stability Board, *2013 Update of Group of Global Systemically Important Banks (G-SIBs)*, 11 November, 2013. Available online at: http://www.financialstabilityboard.org/2013/11/r_131111/.

Hautsch, N., Schaumburg, J. and Schienle, M., Financial network systemic risk contributions. Technical report, CFS Working Paper, 2013.

Jackson, M.O. and Rogers, B.W., Meeting strangers and friends of friends: How random are social networks? *Amer. Econ. Rev.*, 2007, **97**(3), 890–915.

Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H., Visual analytics: Scope and challenges. In *Visual Data Mining*, edited by S.J., Simoff, M.H., Böhlen and A. Mazeika, pp. 76–90, 2008 (Springer: Heidelberg).

Laeven, L. and Valencia, F., Resolution of banking crises: The good, the bad, and the ugly. International Monetary Fund, 2010. Available online at: https://www.imf.org/external/pubs/ft/wp/2010/wp10146.pdf.

Malkiel, B.G., The efficient market hypothesis and its critics. *J. Econ. Perspect.*, 2003, **17**(1), 59–82.

Malo, P., Sinha, A., Takala, P., Ahlgren, O. and Lappalainen, I., Learning the roles of directional expressions and domain concepts in financial news analysis. In *2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, Dallas, TX, pp. 945–954, 2013 (IEEE).

Mikolov, T., Chen, K., Corrado, G. and Dean, J., Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013. Available online at: https://code.google.com/p/word2vec/.

Newman, M.E.J., Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 2001, **64**(1), 016132.

Opsahl, T., Agneessens, F. and Skvoretz, J., Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc. Networks*, 2010, **32**(3), 245–251.

Özgür, A., Cetin, B. and Bingol, H., Co-occurrence network of reuters news. *Int. J. Modern Phys. C*, 2008, **19**(5), 689–702.

Rönnqvist, S. and Sarlin, P., From text to bank interrelation maps. In *IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, London, 2014.

Sarlin, P., On policymakers' loss functions and the evaluation of early warning systems. *Econ. Lett.*, 2013, **119**(1), 1–7.

Sarlin, P., Macroprudential oversight, risk communication and visualization. LSE SP Working Paper No. 4, 2014.

Soramäki, K., Bech, M.L., Arnold, J., Glass, R.J. and Beyeler, W.E., The topology of interbank payment flows. *Physica A*, 2007, **379**(1), 317–333.

Stephenson, K. and Zelen, M., Rethinking centrality: Methods and examples. *Soc. Networks*, 1989, **11**(1), 1–37.

Turney, P.D. and Pantel, P., From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 2010, **37**(1), 141–188.

Watts, D.J. and Strogatz, S.H., Collective dynamics of 'small-world' networks. *Nature*, 1998, **393**(6684), 440–442.

Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V. and Garner, H.R., Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 2004, **20**(3), 389–398.

Appendix 1. Data

Table A1. A list of banks and their labels.

European LCBG and G-SIB		European LCBG	
Label	Name	Label	Name
Agricole	Credit Agricole Groupe	ABN-AMRO	ABN AMRO Bank NV
BBVA	Banco Bilbao Vizcaya Argenta	Bankia	Bankia SA
BPCE	Groupe BPCE	Commerzbank	Commerzbank AG
BNP	BNP Paribas	CreditMutuel	Credit Mutuel Group
Barclays	Barclays PLC	DZBank	DZ Bank AG
CreditSuisse	Credit Suisse Group AG	Danske	Danske Bank A/S
Deutsche	Deutsche Bank AG	Intesa	Intesa Sanpaolo
HSBC	HSBC Holdings PLC	LaCaixa	La Caixa
ING	ING Bank NV	LandesbankBW	Landesbank Baden-Württemberg
Nordea	Nordea Bank AB	Lloyds	Lloyds Banking Group PLC
RBS	Royal Bank of Scotland	Rabobank	Rabobank Group
Santander	Banco Santander SA		
SocGen	Group Societe Generale SA		
StanChart	Standard Chartered PLC		
UBS	UBS AG		

Appendix 2. Usefulness of early-warning models

Early-warning models require evaluation criteria that account for the nature of low-probability, high-impact events. Following Sarlin (2013), the signal evaluation framework focuses on a Policy-maker with relative preferences between type I and II errors, and the usefulness that she derives using a model, in relation to not using it. To mimic an ideal leading indicator, we build a binary state variable $C_j(h) \in \{0, 1\}$ for observation j (where $j = 1, 2, \ldots, N$) given a specified forecast horizon h. Let $C_j(h)$ be a binary indicator that is one during pre-crisis periods and zero otherwise. For detecting events C_j using information from indicators, we estimate the probability of a crisis occurrence $p_j \in [0, 1]$, for which we use herein logistic regression. The probability p_j is turned into a binary prediction P_j, which takes the value one if p_j exceeds a specified threshold $\lambda \in [0, 1]$ and zero otherwise. The correspondence between the prediction P_j and the ideal leading indicator C_j can then be summarized into a so-called contingency matrix.

The frequencies of prediction–realization combinations in the contingency matrix are used for computing a wide range of quantitative measures of classification performance. Beyond measures of overall accuracy, a policy-maker can be thought to be primarily concerned with two types of errors: issuing a false alarm and missing a crisis. The evaluation framework described below is based upon that in Sarlin (2013) for turning policy-makers' preferences into a loss function, where the policy-maker has relative preferences between type I and II errors. While type I errors represent the share of missed crises to the frequency of crises $T_1 \in [0, 1] = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN})$, type II errors represent the share of issued false alarms to the frequency of tranquil periods $T_2 \in [0, 1] = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$. Given probabilities p_j of a model, the policy-maker then optimizes the threshold λ such that her loss is minimized. The loss of a policy-maker includes T_1 and T_2, weighted by relative preferences between missing crises (μ) and issuing false alarms ($1 - \mu$). By accounting for unconditional probabilities of crises $P_1 = P(C = 1)$ and tranquil periods $P_2 = P(C = 0) = 1 - P_1$, the loss function can be written as follows:

$$L(\mu) = \mu T_1 P_1 + (1 - \mu) T_2 P_2 \tag{B1}$$

Table B1. A contingency matrix.

		Actual class C_j	
		Crisis	No crisis
Predicted class P_j	Signal	Correct call *True positive (TP)*	False alarm *False positive (FP)*
	No signal	Missed crisis *False negative (FN)*	Correct silence *True negative (TN)*

where $\mu \in [0, 1]$ represents the relative preferences of missing crises and $1 - \mu$ of giving false alarms, T_1 the type I errors, and T_2 the type II errors. P_1 refers to the size of the crisis class and P_2 to the size of the tranquil class. Further, the usefulness of a model can be defined in a more intuitive manner. First, the absolute usefulness (U_a) is given by:

$$U_a(\mu) = \min(\mu P_1, (1 - \mu) P_2) - L(\mu), \tag{B2}$$

which computes the superiority of a model in relation to not using any model. As the unconditional probabilities are commonly unbalanced and the policy-maker may be more concerned about the rare class, a policy-maker could achieve a loss of $\min(\mu P_1, (1 - \mu) P_2)$ by either always or never signalling a crisis. This predicament highlights the challenge in building a useful early-warning model: with an imperfect model, it would otherwise easily pay off for the policy-maker to always signal the high-frequency class.

Second, we can compute the relative usefulness U_r as follows:

$$U_r(\mu) = \frac{U_a(\mu)}{\min(\mu P_1, (1 - \mu) P_2)}, \tag{B3}$$

where U_a of the model is compared with the maximum possible usefulness of the model. That is, the loss of disregarding the model is the maximum available usefulness. Hence, U_r reports U_a as a share of the usefulness that a policy-maker would gain with a perfectly performing model, which supports interpretation of the measure.

II

Paper II

Bank distress in the news: Describing events through deep learning

S. Rönnqvist and P. Sarlin (2017). *Neurocomputing*, 264(1):57–70

Neurocomputing 264 (2017) 57–70

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Bank distress in the news: Describing events through deep learning

Samuel Rönnqvist[a,b,*], Peter Sarlin[c,d]

[a] Turku Centre for Computer Science – TUCS, Department of Information Technologies, Åbo Akademi University, Turku, Finland
[b] Applied Computational Linguistics Lab, Goethe University Frankfurt am Main, Germany
[c] Department of Economics, Hanken School of Economics, Helsinki, Finland
[d] RiskLab Finland, Arcada University of Applied Sciences, Helsinki, Finland

ARTICLE INFO

Article history:
Received 24 March 2016
Revised 26 October 2016
Accepted 6 December 2016
Available online 16 June 2017

Keywords:
Neural networks
Text mining
Event detection
Bank distress
Distributional semantics
Financial risk

ABSTRACT

While many models are purposed for detecting the occurrence of significant events in financial systems, the task of providing qualitative detail on the developments is not usually as well automated. We present a deep learning approach for detecting relevant discussion in text and extracting natural language descriptions of events. Supervised by only a small set of event information, comprising entity names and dates, the model is leveraged by unsupervised learning of semantic vector representations on extensive text data. We demonstrate applicability to the study of financial risk based on news (6.6M articles), particularly bank distress and government interventions (243 events), where indices can signal the level of bank-stress-related reporting at the entity level, or aggregated at national or European level, while being coupled with explanations. Thus, we exemplify how text, as timely, widely available and descriptive data, can serve as a useful complementary source of information for financial and systemic risk analytics.

1. Introduction

Text analytics presents both major opportunities and challenges. On the one hand, text data is rich in information and can be harnessed in traditional ways such as for prediction tasks, while its descriptive depth also supports qualitative and exploratory, yet highly data-driven, analysis. On the other hand, decoding and utilizing the expressive detail of human language is prohibitively difficult. In computational terms, text consists of high-dimensional and often ambiguous symbolic input (words), the semantics of which is a product of complex interactions between parts of the sequences in which they occur (phrases, sentences, paragraphs, etc.). Text is referred to as sparse data due to the high variability relative to number of samples, and unstructured data as the underlying linguistic structure must be inferred from the surface form as part of the analysis process.

We recognize that many applications of text analytics use linguistically rather naïve methods, typically operating on a bag-of-words assumption, disregarding word order and operating at the symbolic word-level alone. While these applications generally constitute pioneering work in their respective areas, there is currently ample opportunity for advancement, in particular in the intersection between machine learning, computational linguistics and economics. Following the deep learning paradigm, recent developments in natural language processing [37] open up for highly data-driven but linguistically more accommodating analysis methods based on semantic representation learning, which easily can be applied to new domains and tasks.

In this paper, we propose a deep learning setup to address the challenge of building a predictive model able to detect infrequent, coinciding events based on the sparse and unstructured input of text, while leveraging the text data to describe the events as well. Our method includes a heuristic to label text by event information, unsupervised semantic modeling, predictive modeling, aggregation of the prediction signals into indices, and the eventual extraction of descriptions. The approach is to our knowledge novel in providing text descriptions of events defined by non-descriptive data. We show how it can be applied to the study of risks in the financial system, with relatively little effort required in terms of collecting data for supervision in new tasks, which can be a prohibitive aspect of text analytics.

The study of bank distress is a prime example of a field where the use of text data remains largely uncharted, typically lacking both customized linguistic resources and clear goals for how to best utilize text, which motivates the focus on adaptive methods. Supervised by only a small set of bank distress events we demonstrate that the method we put forward can provide an index over coinciding stress-related reporting in news over time, which we then use to automatically retrieve descriptions of the events. We expect the method accordingly to be applicable to any type of

* Corresponding author at: Turku Centre for Computer Science – TUCS, Department of Information Technologies, Åbo Akademi University, Turku, Finland
E-mail address: sronnqvi@abo.fi (S. Rönnqvist).

http://dx.doi.org/10.1016/j.neucom.2016.12.110
0925-2312/

event that recurringly figures in text over time, in connection to specific entities.

In the following section, we discuss previous work related to the problem setting and work that has utilized text data for similar tasks. Deep learning background and our setup, including semantic modeling, predictive modeling and evaluation, extraction of descriptions and the related indices are explained in Section 3. Finally, we report our experiments in Section 4, demonstrating the applicability of our approach to the study of bank stress.

2. Related work

The automatic identification of events in chronological text such as news has been explored at least since the 90s, when a DARPA-coordinated effort was organized [1] that set the foundation for what is known as *topic detection and tracking* (TDT), where news streams are analyzed in order to identify reporting on new events as well as recurring reporting relating to earlier events. The early detection and tracking methods were data-driven, based on clustering in particular, and intended to capture any kind of event (see, e.g., [43]).

A related area of research that since has emerged, mainly stemming from the area of information extraction, is *event extraction*, which aims at extracting complex structured information about events in terms of pre-defined types of events and entities, as well as attributes of events and roles of entities (cf., e.g., [7]). The event extraction techniques focus on identifying and extracting more specific types of information, with explicit semantic interpretation, in contrast to the TDT approach. As the information of interest is often particular to an expert domain and task, the techniques tend to require substantial expert guidance in terms of designing linguistic patterns or annotating text, which makes them less applicable in new domains where fewer resources may be available to target specific tasks and the information of interest may be difficult to strictly define. Efforts focusing on the financial domain and identification of specific types of risk include [8,9,16]. Tanev et al. [40] also explore the combination of data-driven preprocessing with the knowledge-driven approach to extracting events, as they monitor violent and disaster events in news. Hogenboom et al. [17] provide a thorough overview of how event extraction has evolved in various fields.

Parallel to this view on event discovery, which naturally places description of events at its heart, non-text data sources have also been investigated for the detection of significant events, or the risk thereof, such as failure of companies using machine learning [2,12]. The focus is then primarily on estimating the likelihood that a particular type of event will occur. While the specification of events in text mining tends to be more idiosyncratic to the input data, the events in distress prediction tend to be specified by when they occur and what entities they involve, as is the case in this paper, too. Such event specifications are easier to recombine with new data, including text data given appropriate modeling.

In particular, prediction of bank distress has been a major topic both before and following the global financial crisis. Many efforts are concerned with identifying the build-up of risk at early stages, often relying upon aggregated accounting data to measure imbalances (e.g., [5,11,23]). Despite their rich information content, accounting data pose major challenges due to restricted access, as well as low reporting frequency and long publication lags. A widely available and more timely source of information is the use of market data to indicate imbalances, stress and volatility (e.g., [13,25]). Yet, market prices provide little or no descriptive information per se, and only yield information about listed companies or companies' traded instruments (such as Credit Default Swaps). This points to the potential value of text as a source for understanding events such as bank distress. More generally, central banks are starting to recognize the utility of text data in financial risk analytics, too [6,18].

The literature on text-based computational methods for measuring risk or distress is still rather scarce and scattered. For instance, Nyman et al. [28] analyze sentiment trends in news narratives in terms of excitement/anxiety and find increased consensus to reflect pre-crisis market exuberance, Soo [38] analyzes the connection between sentiment in news and the housing market and Cerchiello et al. [10] analyse bank risk contagion with both market prices and sentiment index. All three approaches rely on manually-crafted dictionaries of sentiment-bearing words. While such analysis can provide interesting insight as early work on processing expressions in text to study risk, the approach is generally limiting as dictionaries are cumbersome to adapt to specific tasks, incomplete and unable to handle semantics beyond single words well. Nevertheless, sentiment analysis based on such simple approaches works quite well due to the fact that it relies on human emotions as strong priors in a way that generalizes across tasks and data, and because lower recall may be countered by the scale of the data.

Malo et al. [22] explore a linguistically more sophisticated approach that models financial sentiment compositionally, although without semantic generalization, supervised by a custom data set of annotated phrases. Hogenboom et al. [16] integrate their linguistically aware event extraction techniques with the conventional Value at Risk model to account for certain cases of event-driven market effects.

Data-driven approaches, such as Wang and Hua [42] predicting volatility of company stocks from earning calls, may avoid the issues of handcrafted features and manually annotated corpora. Their method, although allegedly providing good predictive performance gains, offers only limited insight into the risk-related language of the underlying text data. It also leaves room for further improvements with regard to the semantic modeling of individual words and sequences of words, which we address. Further, Lischinsky [21] performs a crisis-related discourse analysis of corporate annual reports using standard corpus-linguistic tools, including some data-driven methods that enable exploration based on a few seed words. His analysis focuses extensively on individual words and their qualitative interpretation as part of a crisis discourse, which likewise provides rather limited insight compared to what full sentences are able to communicate. Finally, Rönnqvist and Sarlin [30] construct network models of bank interrelations based on co-occurrence in news, and assess the information centrality of individual banks with regard to the surrounding banking system, a fully data-driven approach that could be further enhanced by semantic modeling and conditioning.

In the following, we introduce the deep learning approach and our particular model, along with further relevant previous work.

3. Methods

Characterized in part by the deep, many-layered neural networks, a prevailing idea of the deep learning paradigm is that machine learning systems can become more accurate and flexible when we allow for abstract representations of data to be successively learned, rather than handcrafted through classical feature engineering. By modeling the input data before modeling specific tasks, the networks can learn about regularities in the world and generalize over them, which improves performance on supervised task learning. For a recent general survey on deep learning confer Schmidhuber [34], and for a more explicit discussion of deep learning in natural language processing see Socher and Manning [37]. Moreover, Bengio et al. [4] provide a thorough review on the emerging topic of representation learning itself.

While manually designed features help bring structure to the learning task through the knowledge they encode, they often suffer

problems of being over-specified, incomplete and laborious to develop. Especially regarding natural language processing, this limits the robustness of text mining systems and their ability to generalize across languages, domains and tasks. By exploiting statistical properties of the data, features can be learned in an unsupervised fashion instead, which allows for large-scale training not limited by the scarcity of annotated data. Such intensively data-driven, deep learning approaches have in recent years led to numerous breakthroughs in application domains such as computer vision and natural language processing, where a common theme is the use of unsupervised pre-training to effectively support supervised learning of deep networks [34]. We apply the same idea in modeling event-related language in text.

3.1. Labeling text by event data

The modeling is founded on connecting two types of data, text and event data, by entities and chronology. An event data set contains information on dates and names of involved entities, relating to the specific type of event to be modeled. First, a set of regular expression patterns is used to locate the entity names as they occur in the text. Second, an event is associated by the date it occurred and by the relevant timestamp of the document.

In this paper, we focus on news text where publication date is used for matching articles in time, and entity occurrences are indexed at the sentence level. Each sentence s and occurring entity b are cross-referenced against the event data in order to cast the pair as event *coinciding* (1), *non-coinciding* (0) or ambiguous (undefined), according to an inner (W_{in}) and outer (W_{out}) time window. Formally, the label is defined as:

$$e_{s,b} = \begin{cases} 1, & \text{if } d_s - d_e \in W_{in} \\ 0, & \text{if } d_s - d_e \notin W_{out} \end{cases}$$

where for the intervals W holds that $W_{in} \subset W_{out}$. I.e., we label each entity occurrence and its sentence as likely to discuss the event or not likely, whereas uncertain cases that fall outside W_{in} but within W_{out} are not used. Given this heuristic, we effectively expand the data set that is to serve as supervision signal, and the predictive model will learn to generalize across examples and associate relevant language in the text data to the modeled event type.

3.2. Modeling

We are interested in modeling the semantics of words and semantic compositionality of sequences of words to obtain suitable representations of the content of the news, to use as features for predicting events and associating text descriptions. At the word level, distributional semantics exploits the linguistic property that words of similar meaning tend to occur in similar contexts [14]. Word contexts are modeled to yield distributed representations of word semantics as vectors, as opposed to declarative formats, which allow measuring of semantic similarities and detecting analogies without supervision, given substantial amounts of text [24,35,36]. The distributional semantic modeling captures the nature of words in a broader sense, in the directions of syntax and pragmatics. These word vectors provide an embedding into a continuous semantic space where the symbolic input of words can be geometrically related to each other, thus supporting both the predictive modeling in this paper and a multitude of other natural language processing tasks (e.g., tagging, parsing and relation extraction) [4,37].

While traditionally modeled by counting of context words, predictive models have eventually taken the lead in terms of performance [3]. Neural network language models in particular have proved useful for semantic modeling, and are especially practical to incorporate into deep learning setups due to their dense vectors and the unified neural framework for learning. Mikolov et al. [24] have put forward an efficient neural method that can learn highly accurate word vectors as it can train on massive data sets in practical time (a billion words in the order of a day on standard architecture).

Subsequently, Le and Mikolov [20] extended the model in order to represent compositional semantics (cf. [26]) of sequences of words, from sentences to the length of documents, which they demonstrated to provide state-of-the-art performance on sentiment analysis of movie reviews. Methods based on other neural architectures and explicit sentence structure have since gained slightly improved performance [19,39], but require parse trees as pre-structured input and are therefore not as flexible. Analogous to the sentiment analysis task, we employ the distributed memory method of Le and Mikolov to learn vectors for sentences in news articles, where entities are mentioned, and use them for learning to predict the probability of an event. Hence, when providing bank distress events, the task can be understood as a type of risk sentiment analysis that models language specific to the type of event, rather than more general expression of emotions explicitly.

Text sequences are also commonly modeled by recurrent neural networks such as Long Short-Term Memory (LSTM) networks [15], but these are not as efficient as feed-forward topologies with fixed context size in terms of speed. The sentence vector we use is a practical fixed-size representation suitable as input to a feed-forward network. The input sequence of words may have a vocabulary size in the order of a million words, but the sentence vector represents the necessary semantics of each sentence as a single dense vector with a dimensionality of typically 50–1000. The reduction from sparse sequence to a fixed-length, dense representation helps train the predictive model against a signal corresponding to a comparatively tiny number of events.

Our deep neural network for predicting events from text, outlined in Fig. 1, is trained in two steps: through learning of sentence vectors as pre-training (Fig. 1a), followed by supervised learning against the event signal e (Fig. 1b). The use of the distributed memory model of [20] in the first step is explained in the following.

The modeling of word-level semantics works by running a sliding window over text, taking a sequence of words as input and learning to predict the next word (e.g., the 8th in a sequence), using a feed-forward topology where a projection layer in the middle provides the semantic vectors once the connection weights have been learned. A semantic vector V_i is the fixed-length, real-valued pattern of activations reaching the projection layer for network input i. The projection layer provides a linear combination that enables efficient training on large data sets, which is important in achieving accurate semantic vectors. In addition, the procedure of [20] for sentence vector training includes the sentence ID as input, functioning as a memory for the model that allows the vector to capture the semantics of continuous sequences rather than only single words; the sentence ID in fact can be thought of as an extra word representing the sentence as global context and informing the prediction of the next word. While the prediction from word context to word constitutes a basic neural language model, the sentence ID conditions the model on the sentence and forces the sentence vector to capture the semantics that is particular to the sentence rather than the language overall. Formally, the pre-training step seeks to maximize the average log probability:

$$\frac{1}{k-n}\sum_{i=1}^{k-n} \log p(w_{i+n}|s, w_i, \ldots, w_{i+n-1})$$

over the sequence of training words $w_1, w_2, \ldots, w_k$ in sentence s with word context size n. In the neural network, an efficient binary

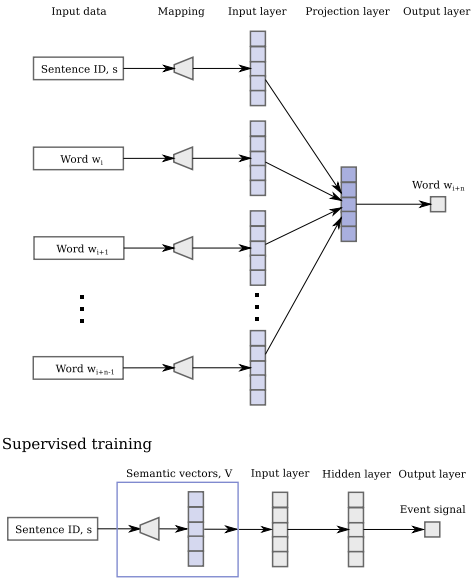


Fig. 1. Deep neural network setup for (a) pre-training of semantic vectors, and (b) supervised training against event signal e.

Huffman tree coding is used to map sentence IDs and words to activation patterns in the input layer and the hierarchical softmax output layer (by referencing vectors of a matrix D), which imposes a basic organization of words by frequency. The projection layer output is a function of the average of sentence vector V_s and word vectors of the context $\{V_{w_j} | j \in [i, i+n[\}$, which means that a single vector can easily be extracted once the model is trained. The sentence vector is extractable as $V_s = \beta + UD_s$, where U is the learned projection layer weight matrix and β is the bias parameter.

The second modeling step (Fig. 1b) is a normal feed-forward network fed by the sentence vectors V_s (pertaining to the set of sentences S), which we train by Nesterov's Accelerated Gradient [27] and backpropagation [32] to predict distress events $e \in \{0, 1\}$. Hence, the objective is to maximize the average log probability:

$$\frac{1}{|S|} \sum_{s \in S} \log p(e_s | V_s)$$

The network has two output nodes for $e \in \{0, 1\}$ in a softmax layer that applies a cross-entropy loss function. In the trained network, the posterior probability $M(V_s) = p(e_s = 1 | V_s)$ reflects the relevance of sentence s to the modeled event type and is derived by:

$$p(e_s = j | V_s) = \frac{e^{y_j}}{e^{y_0} + e^{y_1}}; \; y = \sigma(\beta^2 + U^2 \sigma(\beta^1 + U^1 V_s))$$

where σ can be any non-linear activation function (e.g., sigmoid, hyperbolic tangent or rectified linear) and U are again the learned weight matrices.

In the following sections, we discuss how the model is used for classification and evaluated by its classification performance, as we apply a threshold on the model output $M(V_s)$, as well as on aggregate functions of it.

3.3. Evaluation and aggregation

Assuming that the distribution of events for a particular entity is sparse over time, the procedure for matching events to text produces examples with skewed class frequencies. Moreover, it is likely that the user has an imbalanced preference between types of errors, preferring a sensitive system to detect possible events and provide means for further investigation in the form of descriptions, rather than missing an event. This requires extra care in evaluation.

We evaluate the performance of the predictive model to guide hyperparameter optimization and asses the quality of indices that it will produce, and importantly to provide a quantitative quality assurance for the information content of the descriptions we extract. We use the relative Usefulness measure (U_r) by Sarlin [33], as it is commonly used in distress prediction and intuitively incorporates both error type preference (μ) and relative performance gain of the model over consistently choosing the majority class. Based on the combination of negative/positive observations (obs $\in \{0, 1\}$) and negative/positive predictions (pred $\in \{0, 1\}$), we obtain the cases of true negative ($TN \equiv obs = 0 \wedge pred = 0$), false negative ($FN \equiv obs = 1 \wedge pred = 0$), false positive ($FP \equiv obs = 0 \wedge pred = 1$) and true positive ($TP \equiv obs = 1 \wedge pred = 1$), for which we can estimate probabilities when evaluating our predictive model. Further, we define the baseline loss L_b to be the best guess according to prior probabilities $p(\text{obs})$ and error preferences μ (Eq. (1)) and the model loss L_m (Eq. (2)):

$$L_b = \min \begin{cases} \mu \cdot p(\text{obs} = 1) \\ (1 - \mu) \cdot p(\text{obs} = 0) \end{cases} \tag{1}$$

$$L_m = \mu \cdot p(FN) + (1 - \mu) \cdot p(FP) \tag{2}$$

From the loss functions we derive Usefulness in absolute (U_a) and relative terms (U_r):

$$U_r = \frac{U_a}{L_b} = \frac{L_b - L_m}{L_b} \tag{3}$$

While absolute Usefulness U_a measures the gain vis-à-vis the baseline case, relative Usefulness U_r relates gain to that of a perfect model (i.e., Eq. (5) with $L_m = 0 \Rightarrow U_a = L_b$). Usefulness functions both as a proxy for benchmarking the model (testing) and to optimize its hyperparameters (validation). Usefulness can also be related to the in text mining widely used F-score [41] (based on precision $= p(\text{obs} = 1 | \text{pred} = 1)$ and recall $= p(\text{pred} = 1 | \text{obs} = 1)$):

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{4}$$

which similarly can account for varying preferences by its β parameter, although not gain. The F_β-score assigns β times as much importance to recall as to precision (i.e., preference for completeness over exactness) [41], which is analogous to but not directly transferable to the μ parameter in the Usefulness measure. While the F-score is commonly seen to maximize completeness versus exactness of true positives, the parameter can also be seen as a priority to minimize false negatives versus false positives (FN prioritized over FP when $\beta > 1$). As a heuristic, we map the balanced, standard F_1-score with $\beta = 1$ to U_r with $\mu = 0.5$, and match deviations from these preferences according to $\beta = \mu/(1 - \mu)$.

In order to influence the sensitivity of the model, we may classify a sentence by a threshold on the positive-class posterior probability:

$$p(e_s = 1 | V_s) \geq t$$

The threshold is optimized on the validation set with respect to Usefulness at a given preference, and applied to the test set for evaluation.

However, evaluating classification at an aggregated entity level rather than the level of sentence instances is more suitable to the use case, and likely more robust as the classification then combines evidence from multiple observed occurrences in the text. Instead of the direct posterior probability, at the entity level we classify by the index defined in Eq. (5) below; i.e., an event is signaled for the entity if:

$$I(p, b) \geq t$$

Furthermore, evaluation on the sentence vector level with a randomized set split into train, validation and test set may produce somewhat optimistic results, as specific language related to one particular event can be expected to be shared among several instances. Thus, the evaluation would not truly reflect how well the model can be expected to generalize across events of the same type, including future occurrences. To counter the bias, we sample the cross-validation folds according to a *leave-N-entities-out* strategy (or leave-N-banks-out), based on entity rather than sentence instance, such that discussion about a particular entity is compartmentalized into a single set. In case of very frequent entities that would cause very skewed fold sizes, the instances may be split by period such that the more recent occurrences are placed in the latter set (e.g., test rather than validation set) to minimize possible cross-contamination.

3.4. Event indices

By aggregating posterior probabilities we form an index to reflect the level of event-related reporting about an entity over time, thereby guiding exploration and extraction of descriptions, while it also serves as the signal that we evaluate against. The entity-level relevance index I: $p \times b \rightarrow [0, 1]$ is formalized as:

$$I(p, b) = \frac{1}{|S_{p,b}|} \sum_{s \in S_{p,b}} M(V_s) \tag{5}$$

over the sentences $S_{p,b}$ that mention entity b in period p, where $M(V_s) = p(e_s = 1|V_s)$ gives the posterior probability of the trained neural network model.

In order to obtain better overview, it is motivated to further group entities and aggregate their indices. In the experiments, we first aggregate from sentences to banks, and then from banks to countries to highlight national differences across Europe. The second-level index (or country-level index) is a weighted average, defined as:

$$I'(p, c) = \frac{1}{|B_c|} \sum_{b \in B_c} I(p, b) \cdot |S_{p,b}| \tag{6}$$

where B_c is the set of entities in category/country c. Finally, we define a top-level index that summarizes the level of relevant reporting for all modeled entities as a global average of vectors:

$$I''(p) = \frac{1}{|S_p|} \sum_{s \in S_p} M(V_s) \tag{7}$$

where S_p is the set of vectors for all entity-mentioning sentences in period p.

3.5. Extraction of descriptions

As the neural network in the second step of the setup has been trained and the hyperparameters optimized by cross-validation, it can be applied to sentence vectors V in order to use the posterior probability $M(V)$ as a relevance score with respect to the event type. The indices (Eqs. (5)–(7)) provide overview over time and can highlight peaks and periods with elevated volumes of event-related discussion, which can be more closely investigated by retrieving descriptions of the underlying events.

Given a specific period and entity or set of entities, the basic principle in retrieving descriptions is to filter and rank pieces of text based on the posterior probability of the predictive model for the corresponding semantic vector. In the current setup, we perform the semantic modeling on the sentence level, which simplifies the process of retrieving relevant and specific passages. The semantic modeling can be applied to any type of textual unit, including complete documents, but that requires additional measures for locating the interesting parts within the broader context. Rönnqvist and Sarlin [31] explore this by similarly training a predictive model on document vectors and successfully applying it on word vectors, to weight the relevance individual words within the context. In current experiments, we find that, while their method works for document vectors that are trained on a larger number of words per vector, it does not work as well for sentence vectors, as they tend to be less similar to the word vectors of the same model. Overall, the extracts as presented in Section 4 are qualitatively better when produced based on sentence vectors.

Vectors are trained only for sentences that mention target entity names, as it would be infeasible in terms of memory to model each sentence separately for a large corpus, and because the direct discussion about the entities is of primary interest. The near context of such sentences however tend to support interpretation and are useful to include in presentation. The semantic model supports inference of vectors for at train-time unseen sentences, although with noisier results. We infer vectors and predict the relevance of the sentences immediately before and after sentences in which entities occur, as there is strong dependency between neighboring sentences and a combined score of the expanded context may produce more robust predictions. The combined score for an excerpt is calculated as:

$$x_i = max \begin{cases} M\left(V_{S_i}\right) \\ M\left(\frac{1}{n} \sum_{j=1}^{n} V'_{S_{i-1}}\right) \\ M\left(\frac{1}{n} \sum_{j=1}^{n} V'_{S_{i+1}}\right) \end{cases} \tag{8}$$

which includes one sentence before and after sentence S_i. V' is a stochastic, inferred vector and n is the number of samples (e.g. 100).

The excerpts are ranked according to the score for presentation and offer a preview of the most prominent event-related discussion, which may be retrieved in full from the individual articles. The experiments that follow demonstrate the utility of the excerpts in highlighting the specific forces that drives the index, as we apply the method to model bank distress.

4. Experiments

We test the deep neural network setup for modeling event-related language on European bank distress events and news data, in order to demonstrate the value it can bring in helping to identify and understand past, ongoing or mounting events. In the following, we discuss the data we use, the modeling in practice, and our quantitative evaluation results. Finally, we provide a qualitative analysis of the indices and related events by means of their associated descriptions, going from the general, higher-level view to the more specific.

4.1. Data

The event data set for this study covers data on large European banks as entities, spanning periods before, during and after the global financial crisis of 2007–2009. We include 101 banks for which 243 distress events have been observed during 2007Q3–2012Q2. Following Betz et al. [5], the events include government interventions and state aid, as well as direct failures and distressed

mergers. In addition, we map each bank to the country or countries where it is registered, to allow for aggregation of results to the country level.

The text data consist of news articles from Reuters online archive from the years 2007 to 2014 (Q3). The data set includes 6.6M articles (3.4B words). Bank name occurrences are located using a set of regular expressions that cover common spelling variations and abbreviations. The patterns have been iteratively developed against the data to increase accuracy, with the priority of avoiding false positives (in accordance to [30]). Scanning the corpus, 262k articles and 716k sentences are found to mention any of the 101 target banks.

We set the inner time window from 8 days before to 45 days after the event ($W_{in} = [-8, 45]$), and the outer from 120 days before to 120 days after ($W_{out} = [-120, 120]$), as optimized through the evaluation scheme discussed in Section 3.3. In total, 386k sentences are successfully labled and used for training and evaluation, as they fall within the span of the event data and are not deemed ambiguous cases. As expected, the class distribution is highly skewed, with 9.0% of the 386k cases being labeled as coinciding.

4.2. Semantic pre-training

First, the semantic pre-training step is performed to obtain sentence vectors for each of the 716k sentences, to be used both for training, evaluation and deployment of the model. In order to improve the word representations of the model, by extending the data coverage and letting them capture the semantics of both general English in news reporting as well as bank-specific language, the rest of the corpus is also sampled. This is achieved by running the model without the sentence-ID-related component for sentences without bank occurrences. The whole training process is repeated in multiple iterations with decreasing learning rate. We optimized the sentence vector length to 600 and context size to 5 by cross-validation. We also tested the influence of text sequence lengths, and found that training a vector on multiple sentences achieved slightly worse predictive performance, while vectors trained at sentence and document level were comparable.

4.3. Predictive modeling and evaluation

Following the semantic pre-training, we train a predictive neural network model with 3 layers. The input layer has 600 nodes, corresponding to the semantic vectors, and the output layer has two nodes corresponding to distress/tranquil states. A set of tuples including sentence vectors V_s, entity b and labels $e_{s,\,b}$ are complied as data for modeling.

We evaluate the predictive model with the four combinations of sampling method and level of evaluation, discussed in Section 3.3. The baseline evaluation with random sampling at the level of sentence vectors is reported in Table 1 (left), providing 27.5% relative Usefulness, i.e., performing significantly better than majority class prediction even with the highly skewed class distributions. By comparison, evaluation at the aggregated bank level (classifying by $I(p, b)$ (Eq. (5)) rather than $M(V)$) reduces noise from single sentences and stabilizes prediction, thereby increasing performance to 32.6% (Table 1, center). These results show that the model is effective in linking the relevant pieces of text to the bank distress events, hence, providing a first assurance of the quality of the descriptions we will retrieve. Further, we evaluate based on leave-N-banks-out sampling, i.e., the cross-validation folds of vectors are organized by bank, such that the vectors of banks used for testing are held out of training. While this produces lower Usefulness scores, it is a more realistic estimate of future performance in the context of deploying the model on unseen banks or future data. With vector-level evaluation we reach 8.3% relative Usefulness (Table 1, right), while bank-level aggregation again stabilizes prediction and improves performance to 12.3% of available Usefulness (Table 2).

Table 1
Cross-validated predictive performance as relative Usefulness over preferences between types of error (μ), evaluated at vector and aggregated bank level with random sampling, and at vector level with leave-N-banks-out sampling.

Random sampling					Leave-N-banks-out	
	Vector-level		Aggregated		Vector-level	
μ	$\bar{U}_r(\mu)$	σ_U	$\bar{U}_r(\mu)$	σ_U	$\bar{U}_r(\mu)$	σ_U
0.1	−0.004	0.004	−0.022	0.029	−0.013	0.013
0.3	−0.007	0.004	−0.015	0.013	−0.032	0.026
0.5	0.002	0.005	−0.014	0.010	−0.039	0.036
0.6	0.013	0.007	−0.015	0.012	−0.038	0.039
0.7	0.038	0.011	0.027	0.030	−0.026	0.029
0.8	0.095	0.019	0.156	0.029	−0.008	0.044
0.85	0.157	0.026	0.260	0.030	0.025	0.048
0.875	0.207	0.028	**0.326**	0.030	0.039	0.133
0.9	**0.275**	0.054	0.268	0.031	**0.083**	0.114
0.925	0.253	0.041	0.148	0.040	0.040	0.109
0.95	0.106	0.044	−0.009	0.038	−0.052	0.153

Table 2
Cross-validated predictive performance as relative Usefulness and F-score over preferences between types of error (μ) and recall/precision (β), evaluated at bank level with leave-N-banks-out sampling. Mean confusion matrix values are included, too.

Leave-N-banks-out, aggregated								
μ	$\bar{U}_r(\mu)$	σ_U	$\bar{F}_\beta$	σ_F	$\bar{TN}$	$\bar{FN}$	$\bar{FP}$	$\bar{TP}$
0.1	−0.014	0.042	0.497	0.000	516	68	0	0
0.3	−0.011	0.022	0.087	0.015	516	68	0	0
0.5	−0.015	0.029	0.031	0.013	516	68	0	0
0.6	−0.013	0.027	0.032	0.020	515	68	1	0
0.7	−0.003	0.038	0.087	0.063	511	65	4	3
0.8	0.048	0.154	0.314	0.171	472	53	44	15
0.85	0.122	0.147	0.434	0.153	435	45	80	22
0.875	**0.123**	0.173	0.529	0.174	374	38	142	30
0.9	0.081	0.162	0.629	0.189	308	31	208	37
0.925	−0.006	0.173	0.741	0.190	151	14	364	54
0.95	−0.075	0.160	0.901	0.125	38	4	477	64

We find the optimal network (50 rectified linear hidden nodes), hyperparameters for the NAG training algorithm to train its weights, and threshold on $M(V)$ or $I(b, p)$ for classifying $e \in \{0, 1\}$, after which we evaluate performance by U_r of the optimal model. We trained the network by randomized 5-fold cross validation with one fold for validation and one for testing, in multiple reshuffles of the data set. The evaluation yielded an area under the ROC curve of 0.712 with a standard deviation $\sigma = 0.008$ with random sampling evaluated at vector level, and an area of 0.645 ($\sigma = 0.083$) with leave-N-banks-out sampling evaluated at the aggregated bank level.

Following previous studies [5,29], we make use of a skewed preference $\mu \approx 0.9$ (i.e., missing a crisis is about 9 times worse than falsely signaling one). From the viewpoint of policy, highly skewed preferences are particularly motivated when a signal leads to an internal investigation, and reputation loss or other political effects of false alarms need not be accounted for. While our model is not robust to low levels of μ, we can see in Table 2 that Usefulness is positive and peaking as μ nears 0.9. Meanwhile, F-score is reaching its maximum at the extreme preference, which is an indication of its failure to capture gain over the majority class baseline.

We conclude that at $\mu = 0.9$ with vector-level evaluation and at $\mu = 0.875$ with aggregated evaluation the model has decent predictive performance by capturing up to 33% of available Usefulness and 12% in the more conservative leave-N-banks-out sampled exercise. To relate the results we may confer Betz et al. [5] who obtain U_r of 19-42% and Peltonen et al. [29] with 58-64%. The latter

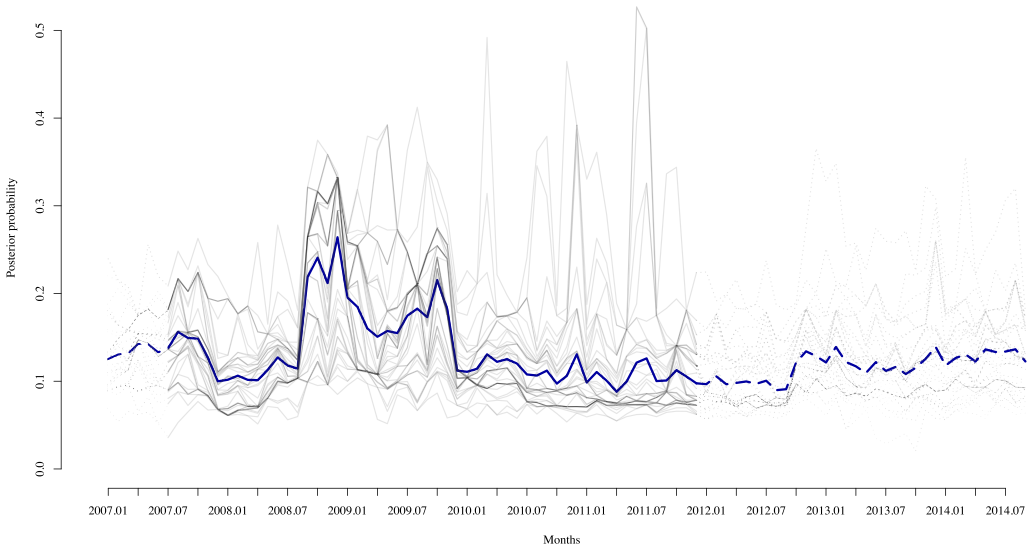


Fig. 2. Raw distress reporting. Distribution of posterior probabilities over time for sentence vectors, indicating the levels of news reporting relating to bank stress. The blue (thick) line indicates mean, faded lines every 2nd percentile, and dotted lines predictions outside the event sample.

incorporates network linkages, which we currently do not model, although this is possible to extract from text as well (cf. [30]). In both cases they test a selection of models to predict bank distress using conventional data sources. These are the most similar experiments available, although not necessarily strictly comparable. A direct comparison of usefulness is in principle impossible as different data and prediction tasks will yield different results, such as the broader sample and earlier forecast horizons in Betz et al. [5]. Nevertheless, our evaluation results show that we are able to extract a stress signal from text alone. While it does not surpass the performance achieved for other tasks and samples, it does achieve acceptable levels and provides a quantitative quality assurance of the text extracts. The results also point toward the likely benefit of incorporating both text and conventional data in bank distress prediction.

4.4. A descriptive stress index for Europe

Having trained the network and evaluated its predictive performance, we can reliably extract indices of stress at the different levels of aggregation together with extracts to describe them. In this section, we discuss patterns recognizable in the top-level view, with a summary of what we are able to learn from the associated descriptions. The following sections continue with a breakdown into countries and banks, which supports a more targeted qualitative analysis.

First, Fig. 2 provides an overview of the raw distress reporting in Europe over the recent years, in terms of distributions of posterior probabilities of the sentence vectors, illustrated through their percentiles. The time span July 2007 to June 2012 is covered by the event data, and the rest is produced by applying the trained model. This distribution communicates the dynamics of the stress situation in Europe, while the mean (index I'' of Eq. (7)) summarizes the general trends.

The index shows a sharp double peak starting September 2008, which coincides with the outbreak of the financial crisis. Prior to the most significant peaks, one can also observe elevated values between August and October 2007, pointing to early discussion on the significance of subprime activities overall and liquidity in European banks. The outbreak of the financial crisis in 2008 is followed by over a year of relatively high stress, where a substantial part of the cross section is elevated. A second significant and similar peak of the stress index is reached in October 2009. At the end of 2010 and 2011, one can observe notable jumps in the most extreme percentiles, whereas the rest of the cross section remains largely unaffected.

At a general level, we observe that the peak in September 2008 relates to overall distress in financial markets due to the collapse of Lehman Brothers in mid-September. However, the fact that values at the top of the distribution appear rather unstable from month to month reflects that different banks are being mentioned over time and usually not persistently across months in distress contexts. By observing increases and peaks in the index of an individual bank or banks in a country, we can identify specific events of possible relevance to distress.

4.5. Country-level stress, descriptions and interpretation

From the general stress index for Europe, this section moves to a more granular perspective on stress, closer to the level of the events being modeled. We measure stress-related discourse for countries for a more targeted stress measure, which also allows for more economic interpretation of developments, as we study the top-ranking excerpts at key points. Thus, we now aggregate posterior probabilities over time for sentence vectors, indicating the levels of news reporting relating to bank stress, but selectively at a country level (according to Eq. (6)). Fig. 3 shows the developments in stress-related discussion for Belgium and Ireland and Fig. 4 for Germany and UK. The figures illustrate stress levels as time series,

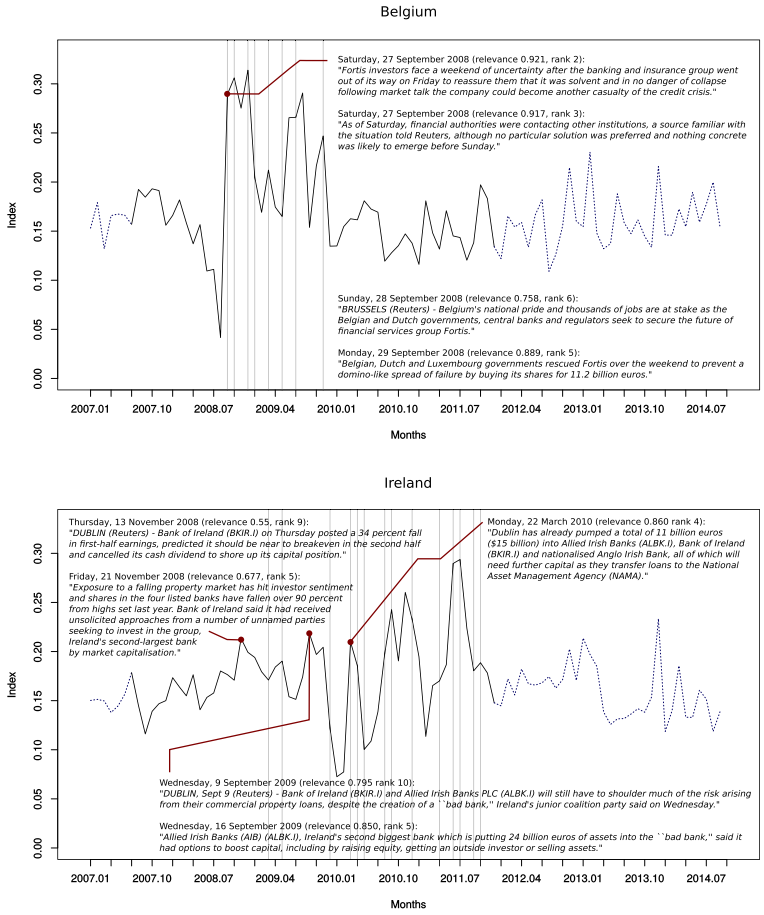


Fig. 3. Distress index for Belgium and Ireland, with key periods marked and informative excerpts selected from the top-10 of each period and country. Vertical lines indicate distress events and dotted lines out-of-sample predictions. Quotes are from Reuters at given dates.

as well as they annotate peaks of distress levels with top-ranked excerpts. In the appendix, we include plots in Figs. A.6 and A.7 for the other countries whose banks we model.

In Fig. 3, the stress levels for Belgium peak in September 2008. Looking at top-ranked excerpts, September 27 is coupled with a range of rumors in media, yet no official release or actions to mitigate the weakened position of particularly Fortis Bank. Then, the next days we see a bailout of Fortis being discussed as the Belgian, Dutch and Luxembourg governments rescued Fortis. Likewise, the lower chart for Ireland in Fig. 3 shows increased concerns over Bank of Ireland and other large Irish banks in November 2008, as both their earnings and shares were significantly falling. After a range of actions by the state, distress levels were still peaking in September 2009, which is particularly related to the amounts that Allied Irish Banks was putting into the Irish "bad bank". Still, in March 2010 three large Irish banks were still transferring large loans to the National Asset Management Agency (NAMA). Thereafter the most acute stress decreased and has since been at lower levels, although remaining somewhat volatile.

Fig. 4 provides similar stress time series and top-ranked excerpts, but for Germany and the UK. Germany can be seen to signal already in August 2007, when IKB's problems were highlighted to potentially lead to "Germany's worst financial crisis in more than 75 years". Three days after this news, Deutsche Bank cut a credit line to IKB, as they were worried about IKB's subprime exposures, which further triggered distress in the German banking sector. One reason to the failure of IKB related to an offshore portfolio that was kept off IKB's balance sheet by Rhineland Funding, which is said to have been explained to the largest shareholder KfW. The same large shareholder is then a few months later involved in helping IKB back on its feet with a hefty 4.8 billion euros, as well as additional smaller support afterwards. For the UK, stress increased in September 2008, relating not only to previous aid to the UK-based Northern Rock but also to Germany's IKB. Here, we see an example

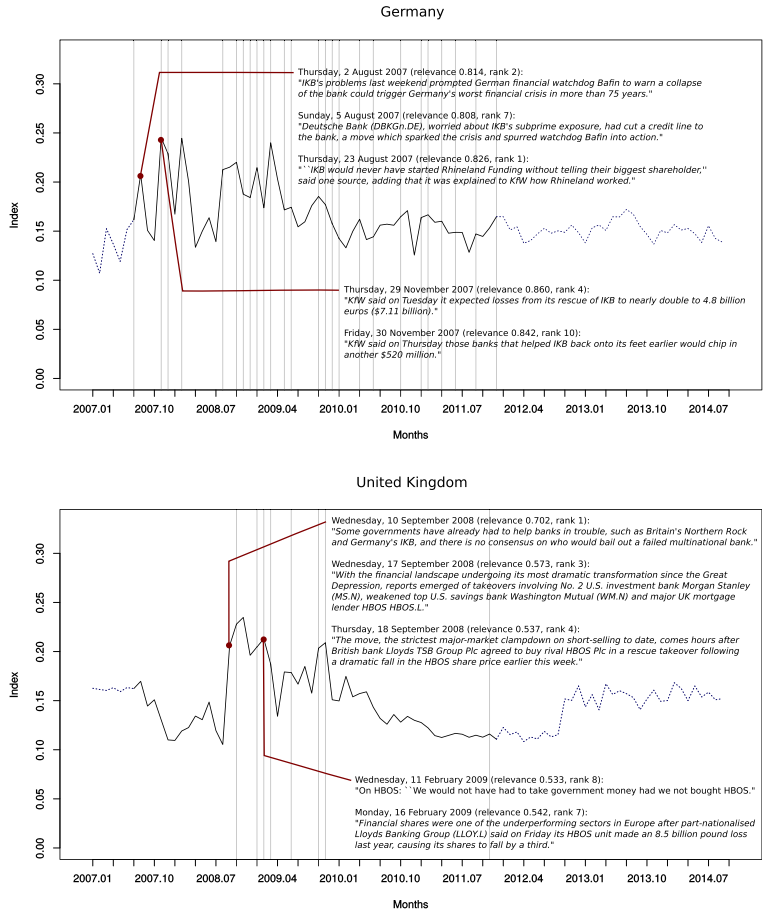


Fig. 4. Distress index for Germany and the United Kingdom, with key periods marked and informative excerpts selected from the top-10 of each period and country. Vertical lines indicate distress events and dotted lines out-of-sample predictions. Quotes are from Reuters at given dates.

of cross-border, systemic effects of bank distress. Only a few days later in conjunction with a strict clampdown on short-selling, UK-based bank Lloyds Group bought rival HBOS in a rescue takeover. Ironically, a few months later in February 2009 Lloyds in partly nationalized as its HBOS unit made an 8.5 billion pounds loss the year before.

4.6. The case of Fortis and IKB bank

This section takes a final step towards more granular output by providing a stress measure for individual banks (according to Eq. (5)). As with the country-level aggregates, we can aggregate posterior probabilities for sentence vectors selectively by bank. This output could be derived for each of the 101 banks, although here we focus on the stress reporting for two banks, namely Fortis and IKB Bank.

One of the early failures among European financial institutions occurred to the Benelux-based Fortis. As was also highlighted in the above described top excerpts for Belgium, Fortis and the rescue procedure was at the core of the discussion as the crisis erupted. We focus on the evolution of the distress index for Fortis, as is shown in Fig. 5. To start with, we can observe that elevated values for the stress index coincide with distress events.

By the first event in September 2008, the index rises to 0.30, which marks the start of a prolonged period of elevated stress. The top-ranked excerpts relate to a range of different issues, such as worries about lacking confidence in the markets and the systemic nature of the unfolding crisis:

> *"Jean-Claude Juncker, also the prime minister of Luxembourg, was asked whether the part nationalisation of Dutch–Belgian bank Fortis FOR. -BR and a new injection of liquidity into money markets by the European Central Bank would restore market confidence. "I can only hope that confidence will come back – financial markets should not forget to take a close look at the health of fundamental data of several banks – and that this casino game, that's*

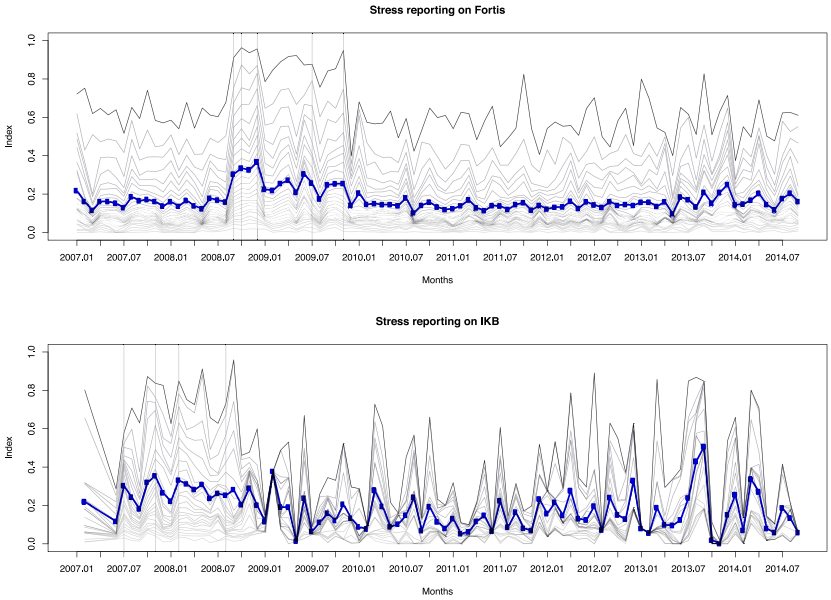


Fig. 5. Indices (blue/thick line) for banks Fortis and IKB indicating the levels of bank stress-related reporting, with faded lines showing every 4th percentile up to the 98th. Vertical lines indicate recorded events.

going on independently from the good fundamentals, stops," he told reporters on the sidelines of a meeting in parliament. Belgian, Dutch and Luxembourg governments rescued Fortis over the weekend to prevent a domino-like spread of failure by buying its shares for 11.2 billion euros."

(Reuters 2008-09-29, relevance 0.963, rank 1)

"Investors also worried if a proposed U.S. rescue would stem the contagion that pushed the British government to takeover troubled mortgage lender Bradford & Bingley BB.L and three European governments to partially nationalize banking and insurance group Fortis FOR.BRFOR.AS."

(Reuters 2008-09-29, relevance 0.923, rank 6)

In October 2008, the top excerpts discuss the continuing developments such as the Benelux governments "carving up" Fortis to sell to private entities, including French BNP Paribas buying control of the arms in Belgium and Luxembourg. Further excerpts highlight the cross-border aspect of the interventions, and the issues it entails:

"The Fortis deal is the biggest cross-border rescue since the full force of the credit crisis swept across the Atlantic into Europe last month, upending banks and rattling saver confidence."

(Reuters 2008-10-06, relevance 0.945, rank 7)

"Dutch Finance Minister Wouter Bos fanned Belgian resentment by telling journalists: 'Many of the problems were hidden in the Belgian part of the Fortis group."

(Reuters 2008-10-05, relevance 0.945, rank 8)

This repeats the message of the already cited news for the UK in September 2008, that "there is no consensus on who would bail out a failed multinational bank", highlighting how the use of text descriptions can provide deeper insight into the multifaceted developments underlying a model signal.

Without a detailed analysis of the discussion around the IKB Bank, we can again conclude from Fig. 5 that the stress index takes high values during the realized events. Generally, the top-ranked discussion herein correlates to a large extent with the early top-ranked discussion for Germany, as was above exemplified. The discussion around the distress events relates to early indications of stress, ties to other German banks and government actions taken during and after the stress episodes. After a period of elevated stress during 2007–2008, the figure illustrates that stress is still fairly volatile and that the most extreme percentiles still take large values. This may relate to the fact that discussion keeps relating to the 2007–2008 distress events, in that the solution to the stress events was an acquisition by a investment company. The private equity firm Lone Star acquired IKB Bank in 2008 with the aim of restructuring and selling the bank, and accordingly any rumors still link it to the original stress discussion during the global financial crisis. Such references to past major stress events may however also be an indication of current concerns about financial stress, thus worth signaling in order to allow further investigation.

5. Conclusions

We have presented a deep-learning-based approach that combines two types of data, news text and basic event information, with the aim of linking the two to describe observed and predicted events. The approach entails unsupervised learning on text in order to model its language and provide semantic vector representations that are used as features for predictive modeling of events. The neural-network-based method that we put forward is able to work with a very small set of events, matched with text through a heuristic, in order to discern what type of language and passages in the text are actually relevant to the modeled event type and phenomenon. The semantic modeling utilizes large amounts of text data to infer abstractions that counter the high variability

and sparsity of language, thus supporting prediction of infrequent events.

The semantic-predictive model can produce indices that indicate the level of relevant discussion over time, overall or related to specific entities or groups thereof. The indices can highlight interesting patterns and offer guidance in the search for relevant events, whereas the model very directly provides means to rank and retrieve pieces of text from news articles in order to describe the quantitative signal.

We demonstrate the usefulness of the method and the possibilities of the approach in general within the study of financial risk, by modeling bank distress events. The indices reflect the level of current reporting related to bank stress over time at multiple levels: for Europe in general, for individual countries and for specific banks. Guided by the indices, users may focus their search and retrieve the relevant reporting of the time, in order to understand the developments regarding, in this case, government interventions and rescues. Our quantitative evaluation of the stress index shows good results and provides an important quality assurance of the descriptions.

The method and our analysis exemplify how text may offer an important complementary source of information for financial and systemic risk analytics, which is readily available, current and rich in descriptive detail. In contrast to traditional information sources, text data offers a possible route to circumvent the issues of privileged access, lagging publication and low granularity, but most importantly does it very directly offer value through the explanatory power of the event-related human language descriptions accompanying the plane signal. We expect the method to be also directly applicable to describe events beyond the financial domain, relating to geopolitics and other significant topics.

We recognize that deep learning approaches are useful in particular to handle the complexities of such new types of data, while offering necessary flexibility when exploring new fields of analysis. Seeking to harness the expressiveness of text, we should continue to look to computational linguistics for support in terms of theoretical foundations and tools.

While we show that it is possible to predict relevance and retrieve informative descriptions of events, we merely scratch the surface of the vast text material in any given cross section with our current method of presentation. A challenge remains in developing methods that are able to meaningfully summarize the broader base that may include a long tail of weakly signaling, subtle expressions. Such signals may be particularly important in order to register and track developments before they materialize in severe and obvious events. Likewise, to really make use of text data as a complement rather than a replacement, traditional sources and text should be integrated in a unified modeling framework in order to achieve the best predictive performance possible, while also keeping the opportunity to explore the descriptions to that signal open.

Acknowledgment

The research has been funded by the Graduate School of Åbo Akademi University and the Turku Centre for Computer Science Graduate Programme. The authors are grateful to Filip Ginter, József Mezei, Tuomas Peltonen and Niko Schenk for their helpful comments. The paper also has benefited from presentation at the Finnish Economic Association XXXVII Annual Meeting (KT-päivat), 12 February 2015, in Helsinki, Finland; the RiskLab/Bank of Finland/European Systemic Risk Board (ESRB) Conference on Systemic Risk Analytics (SRA), 24 September 2015, in Helsinki; the workshop of GI-FG Neuronale Netze and German Neural Networks Society, New Challenges in Neural Computation (NC^2), 10 October 2015, in Aachen, Germany; the IEEE Conference on Computational Intelligence in Financial Engineering and Economics (CIFEr), 9 December 2015, in Cape Town, South Africa; the Financial Stability Seminar at the Riksbank, 12 January 2016, in Stockholm, Sweden; and the 23rd Annual Conference of the Multinational Finance Society, 26 June 2016, in Stockholm, Sweden; as well as from featuring on Bloomberg View. The work presented in this paper has been replicated by Thomson Reuters.

Appendix A.

Figs. A.1 and A.2 provide country-level indices for the countries not included in Figs. 3 and 4, as well as the non-weighted average of all country indices. The individual banks and countries they are mapped to are listed in Table A.1.

Table A1
Target banks and their countries.

Bank	Country	Bank	Country
ABN Amro	NL	Fionia (Nova Bank)	DK
ATE Bank	GR	First Business Bank	GR
Aareal Bank	DE	Fjordbank Mors	DK
Aegon	NL	Fortis Bank	LU, NL, BE
Agricultural Bank of Greece	GR	HBOS	UK
Allied Irish Banks	IE	HSH Nordbank	DE
Alpha Bank	GR	Hellenic	GR
Amagerbanken	DK	Hypo Alpe Adria Group	AT
Anglo Irish Bank	IE	Hypo Real Estate	DE
Attica Bank	GR	Hypo Tirol Bank	AT
BBK	ES	IKB	DE
BNP Paribas	FR	ING	NL
BPCE	FR	Irish Life and Permanent	IE
BPP	PT	Irish Nationwide Building Society	IE
Banca Civica	ES	KBC	BE
Banca Popolare	IT	Kommunalkredit	AT
Banca Popolare di Milano	IT	LBBW	DE
Banco Mare Nostrum	ES	Lloyds TSB	UK
Banco Popolare	IT	Lokken	DK
Banco de Valencia	ES	Magyar Fejlesztesi Bank Zrt	HU
Bank of Cyprus	CY	Marfin Popular Bank	CY
Bank of Ireland	IE	Max Bank	DK
Bankia	ES	Monte dei Paschi di Siena	IT
Banque Populaire	FR	Mortgage and Bank of Latvia	LV
Bawag	AT	National Bank of Greece	GR
BayernLB	DE	NordLB	DE
CAM	ES	Nordea	SE
Caisse d'Epargne	FR	Northern Rock	UK
Caixa General de Depositos	PT	Nova Ljubljanska banka	SI
Caja Castilla-La Mancha	ES	Novacaixagalicia	ES
Caja Espana	ES	OTP Bank Nyrt	HU
Carnegie Investment Bank	SE	Panellinia Bank	GR
Catalunyacaica	ES	Pantebrevsselskabet	DK
Commerzbank	DE	Parex	LV
Cooperative Central Bank	CY	Piraeus Bank	GR, CY
Credit Agricole	FR	Proton Bank	GR
Credit Mutuel	FR	RBS	UK
Credito Valtellinese	IT	RZB Group	AT
Cyprus Development Bank	CY	Roskilde Bank	DK
Cyprus Popular	CY	SNS Reaal	NL
Danske Bank	DK	SachsenLB	DE
Dexia	BE, FR, LU	Societe Generale	FR
Dunfermline	UK	Swedbank	SE
EBH	DK	T-Bank	GR
EBS Building Society	IE	UBS	CH
EFG Eurobank	GR	UNNIM	ES
Eik Bank	DK	USB Bank	CY
Erste Bank	AT	VBAG	AT
Ethias	BE	Vestjysk	DK
FHB Jelzalogbank Nyrt	HU	WestLB	DE
Finansieringsselskabet	DK		

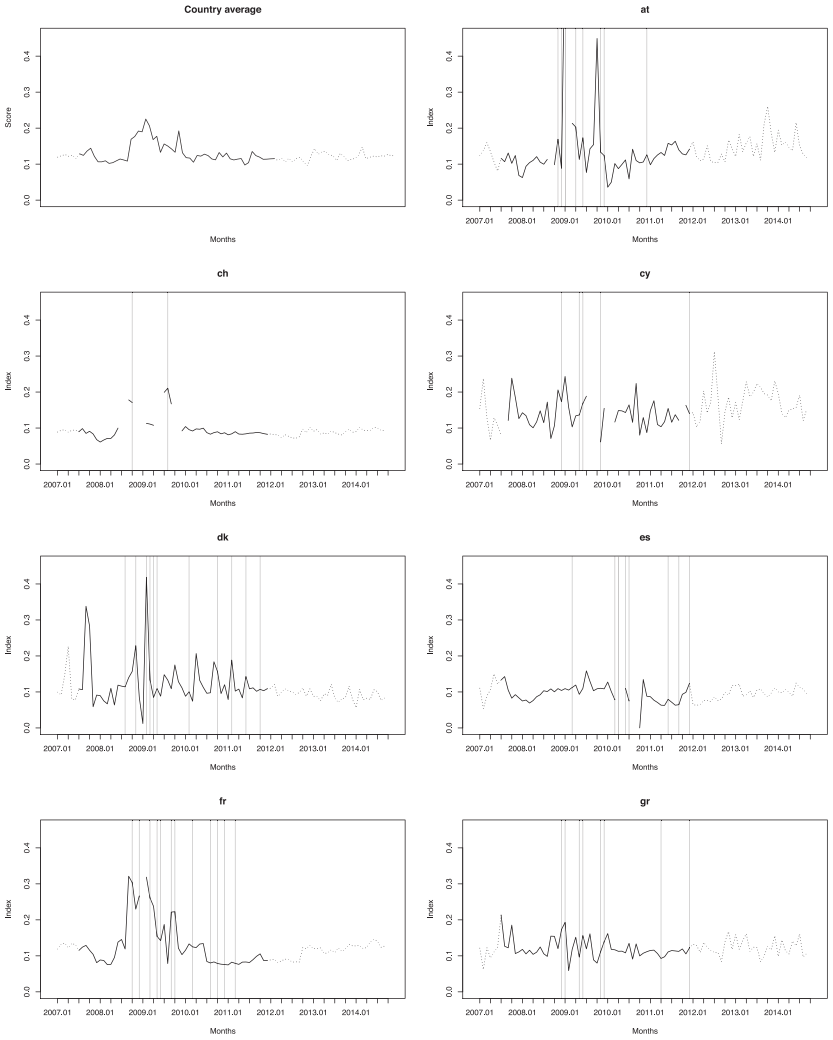


Fig. A1. Distress index for Austria, Switzerland, Cyprus, Denmark, Spain, France, Greece and average of all modeled countries. Vertical lines indicate bank-level distress events and dotted lines out-of-sample predictions.

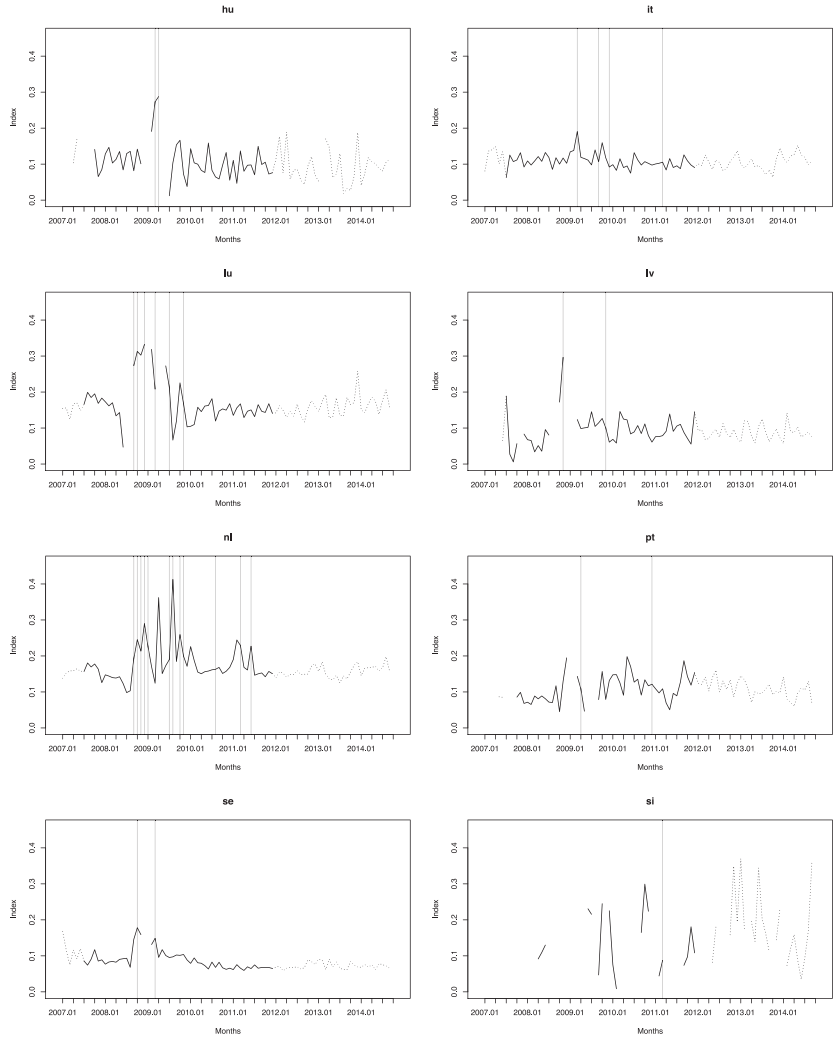


Fig. A2. Distress index for Hungary, Italy, Luxembourg, Latvia, Netherlands, Portugal, Sweden and Slovenia. Vertical lines indicate bank-level distress events and dotted lines out-of-sample predictions.

References

[1] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study final report, in: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] B. Back, T. Laitinen, K. Sere, Neural networks and genetic algorithms for bankruptcy predictions, Expert Syst. Appl. 11 (4) (1996) 407–413, doi:10.1016/S0957-4174(96)00055-3. http://www.sciencedirect.com/science/article/pii/S0957417496000553.

[3] M. Baroni, G. Dinu, G. Kruszewski, Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2014, pp. 238–247.

[4] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intel. 35 (8) (2013) 1798–1828.

[5] F. Betz, S. Oprică, T.A. Peltonen, P. Sarlin, Predicting distress in european banks, J. Bank. Finance 45 (2014) 225–241.

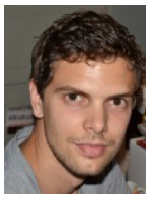
[6] D. Bholat, S. Hansen, P. Santos, C. Schonhardt-Bailey, Text mining for central banks, in: Centre for Central Banking Studies Handbook, vol. 33, Bank of England, 2015.

[7] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, T. Salakoski, Extracting complex biological events with rich graph-based feature sets, in: Proceedings of the BioNLP'09 Shared Task on Event Extraction, 2009, pp. 10–18.

[8] J. Borsje, F. Hogenboom, F. Frasincar, Semi-automatic financial events discovery based on lexico-semantic patterns, Int. J. Web Eng. Technol. 6 (2) (2010) 115–140.

[9] P. Capet, T. Delavallade, T. Nakamura, A. Sandor, C. Tarsitano, S. Voyatzi, A risk assessment system with automatic extraction of event types, in: Proceedings of the International Conference on Intelligent Information Processing, Springer, 2008, pp. 220–229.
[10] P. Cerchiello, P. Giudici, G. Nicola, Big data models of bank risk contagion, in: DEM Working Paper Series, 117, 2016. (02-16).
[11] R.A. Cole, J.W. Gunther, Predicting bank failures: A comparison of on- and off-site monitoring systems, J. Financ. Serv. Res. 13 (1998) 103–117.
[12] A.I. Dimitras, R. Slowinski, R. Susmaga, C. Zopounidis, Business failure prediction using rough sets, Eur. J. Oper. Res. 114 (2) (1999) 263–280, doi:10.1016/S0377-2217(98)00255-0. http://www.sciencedirect.com/science/article/pii/S0377221798002550.
[13] R. Gropp, J. Vesala, G. Vulpes, Equity and bond market signals as leading indicators of bank fragility, J. Money Credit Bank. 38 (2) (2006) 399–428.
[14] Z.S. Harris, Distributional structure, Word 10 (23) (1954) 146–162.
[15] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[16] F. Hogenboom, M. de Winter, F. Frasincar, U. Kaymak, A news event-driven approach for the historical value at risk method, Expert Syst. Appl. 42 (10) (2015) 4667–4675.
[17] F. Hogenboom, F. Frasincar, U. Kaymak, F. de Jong, E. Caron, A survey of event extraction methods from text for decision support systems, Dec. Support Syst. 85 (2016) 12–22, doi:10.1016/j.dss.2016.02.006. http://www.sciencedirect.com/science/article/pii/S0167923616300173.
[18] J. Hokkanen, T. Jacobson, C. Skingsley, M. Tibblin, The Riksbank's future information supply in light of Big Data, in: Economic Commentaries, vol. 17, Sveriges Riksbank, 2015.
[19] O. Irsoy, C. Cardie, Deep recursive neural networks for compositionality in language, in: Advances in Neural Information Processing Systems, 2014, pp. 2096–2104.
[20] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1188–1196.
[21] A. Lischinsky, In times of crisis: a corpus approach to the construction of the global financial crisis in annual reports, Crit. Discourse Stud. 8 (3) (2011) 153–168, doi:10.1080/17405904.2011.589231.
[22] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, P. Takala, Good debt or bad debt: Detecting semantic orientations in economic texts, J. Assoc. Inform. Sci. Technol. 65 (4) (2014) 782–796.
[23] K. Männasoo, D.G. Mayes, Explaining bank distress in Eastern European transition economies, J. Bank. Finance 33 (2009) 244–253.
[24] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Proceedings of Workshop at International Conference on Learning Representations, 2013.
[25] A. Milne, Distance to default and the financial crisis, J. Financ. Stab. 12 (2014) 26–36.
[26] J. Mitchell, M. Lapata, Composition in distributional models of semantics, Cognit. Sci. 34 (8) (2010) 1388–1429.
[27] Y. Nesterov, A method of solving a convex programming problem with convergence rate o (1/k2), in: Soviet Mathematics Doklady, vol. 27, 1983, pp. 372–376.
[28] R. Nyman, D. Gregory, K. Kapadia, P. Ormerod, D. Tuckett, R. Smith, News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment, BoE, Mimeo, 2015.
[29] T. Peltonen, A. Piloui, P. Sarlin, Network Linkages to Predict Bank Distress, ECB Working Paper, No. 1828, 2015.
[30] S. Rönnqvist, P. Sarlin, Bank networks from text: Interrelations, centrality and determinants, Quant. Finance 15 (10) (2015).
[31] S. Rönnqvist, P. Sarlin, Detect & describe: Deep learning of bank stress in the news, in: Proceedings of IEEE Symposium Series on Computational Intelligence, 2015, pp. 890–897, doi:10.1109/SSCI.2015.131.
[32] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
[33] P. Sarlin, On policymakers' loss functions and the evaluation of early warning systems, Econ. Lett. 119 (1) (2013) 1–7.
[34] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015) 85–117.
[35] H. Schütze, Dimensions of meaning, in: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, IEEE Computer Society Press, Los Alamitos, CA, USA, 1992, pp. 787–796. http://dl.acm.org/citation.cfm?id=147877.148132.
[36] H. Schütze, J. Pedersen, Information retrieval based on word senses, in: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, 1995, pp. 161–175.
[37] R. Socher C. Manning. Deep learning for natural language processing (without magic). in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2013). http://nlp.stanford.edu/courses/NAACL2013/.
[38] C.K. Soo, Quantifying Animal Spirits: News Media and Sentiment in the Housing Market, 2013. Ross School of Business Paper No. 1200.
[39] K.S. Tai, R. Socher, C.D. Manning, Improved semantic representations from tree-structured long short-term memory networks, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 26-31, 2015, pp. 1556–1566.
[40] H. Tanev, J. Piskorski, M. Atkinson, Real-Time News Event Extraction for Global Crisis Monitoring, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 207–218. 10.1007/978-3-540-69858-6_21
[41] C.J. Van Rijsbergen, Information Retrieval, 2nd ed., Butterworth, 1979.
[42] W.Y. Wang, Z. Hua, A semiparametric gaussian copula regression model for predicting financial risks from earnings calls, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), 2014.
[43] Y. Yang, J.G. Carbonell, R.D. Brown, T. Pierce, B.T. Archibald, X. Liu, Learning approaches for detecting and tracking news events, IEEE Intel. Syst. 14 (4) (1999) 32–43.

Samuel Rönnqvist is a Ph.D. candidate in Computer Science at Åbo Akademi University (ÅAU; Turku, Finland) and Turku Centre for Computer Science. He has a M.Sc. in Computer Science from ÅAU in 2012. He has been visiting the Technical University of Valencia, Spain (in 2008-09; artificial intelligence, machine learning, bioinformatics), University of Turku (in 2011-12; biomedical text mining) and Goethe University Frankfurt am Main, Germany (in 2015-2016; natural language processing, deep learning). The main focuses of his research are text mining, deep learning and visual analytics, with applications toward, among others, the study of financial, systemic risk.



Peter Sarlin is an Associate Professor of Economics at Hanken School of Economics (Helsinki, Finland), and Director of RiskLab Finland. He is also a visiting scholar with the Center of Excellence SAFE at Goethe University Frankfurt, and a research associate with the Systemic Risk Center at London School of Economics and IWH Halle Institute for Economic Research, as well as a board member of the IEEE Analytics and Risk Technical Committee and the IEEE Computational Finance and Economics Technical Committee. He is an Associate Editor of Journal of Network Theory in Finance and Intelligent Systems in Accounting, Finance & Management. Peter received his Ph.D. (Econ) from the Department of Information Technologies, Åbo Akademi University (Turku, Finland) in 2013. His current research interests include systemic risk, macroprudential supervision, machine learning and visual analytics.

Paper III

Interactive visual exploration of topic models using graphs

S. Rönnqvist and X. Wang and P. Sarlin (2014). In *Proceedings of the Eurographics Conference on Visualization (EuroVis)*, 3 pages. Eurographics Working Group on Data Visualization

Eurographics Conference on Visualization (EuroVis) (2014)
N. Elmqvist, M. Hlawitschka, and J. Kennedy (Editors)

Poster

Interactive Visual Exploration of Topic Models using Graphs

Samuel Rönnqvist[1,2], Xiaolu Wang[2] & Peter Sarlin[3,4]

[1]Turku Centre for Computer Science [2]Åbo Akademi University, Finland [3]Goethe University, Center of Excellence SAFE, Germany
[4]RiskLab at IAMSR, Åbo Akademi University and Arcada University of Applied Sciences

Abstract

Probabilistic topic modeling is a popular and powerful family of tools for uncovering thematic structure in large sets of unstructured text documents. The extensive research into this type of modeling has meet comparatively few studies concerning how to present or visualize topic models in meaningful ways. In this paper, we present a design that uses graphs to visually communicate topic structure and meaning, as uncovered by unsupervised modeling. By connecting topic nodes via descriptive keyterms, the graph representation reveals topic similarities, topic meaning and shared, ambiguous keyterms, while also supporting information retrieval by topic or topic subsets.

Categories and Subject Descriptors (according to ACM CCS): I.5.5 [Pattern Recog.]: Implem.—Interactive systems

1. Introduction

Across domains, we are faced with substantial and often overwhelming amounts of textual data, which present a need for tools to aid in organizing and understanding its content. Text mining techniques are widely used to computationally model human language and extract meaning, and have been successfully applied in many areas, yet, the intricacies of human language constitute an ever-present challenge to computational processing of text. Human involvement is still central to the text mining process [RKPW08], and it emphasizes the necessity for effective visual communication between the computer and the user. The two-way communication enabled by interaction supports a more intimate understanding of the underlying model [War04].

This paper is concerned with probabilistic topic modeling and visual communication of modeling results. Topic modeling algorithms are used to discover latent topics in sets of documents, as a way of uncovering thematic structure. However, visualization of topic models is still a little researched area, which we seek to explore. This answers directly to concerns voiced by one of the authors of the original probabilistic topic modeling algorithm, David M. Blei, who states that "topic models provide new exploratory structure in large collections – how can we best exploit that structure to aid in discovery and exploration?" [Ble12]. He further asserts that interface questions are "essential to topic modeling", but that "making this structure useful requires careful attention to information visualization and [...] user interfaces".

Our contribution is a visualization design that aims to convey the meaning of modeled topics in an intuitive way, as well as how topics relate to each other. It is an interactive graph visualization that connects topics and descriptive keyterms, providing both corpus overview and thematic exploration of documents. A web-based interactive demo is available at: http://risklab.fi/demo/topics/ev14/. Using a corpus of financial patent applications, topics are modeled and presented for exploration, to illustrate the design's utility.

2. Probabilistic Topic Modeling and Visualization

The family of probabilistic topic modeling algorithms has emerged as a widely popular approach to analyzing thematic structure in text. Latent Dirichlet Allocation (LDA) is the most fundamental among them, having inspired a long array of variations [BL09]. Topic modeling is applicable especially to problems where there is little prior knowledge of the content of the text, i.e., where exploratory analysis is needed. LDA infers latent topics in an unsupervised manner, based on word co-occurrence in documents, and provides interpretable output in the form of probabilities. The algorithm takes the desired number of topics as input, assumes that each document may discuss several topics and attempts to identify coherent and meaningful topics by analyzing the terms in each document. By taking the context into account, LDA can help disambiguate single words.

There are two types of relations that are interesting for the

presentation of LDA results: the topic-document relations and the topic-term relations. First, LDA provides a probability distribution over topics for each document, that is, to what degree a document relates to each of the topics. Second, LDA provides topic assignments for each term in a document. The topics provided by LDA are defined by probability distributions over terms. The elementary way of representing the meaning of a topic is through its term probabilities directly, which are based on term frequencies for the topic. Such a representation is rather uninformative to a user, as it gives high ranking to common stop words (e.g., the, a, and) and terms that are general to the whole document corpus (in the case of patents: method, system etc.); better solutions for topic presentation are discussed in Section 3.

Despite the importance carried by visual representation in making topic models useful, the subject has seen limited research effort. Some tools that present topic modeling results rely heavily on text ordered in different fields, but use few visual aids to communicate structure (see, e.g., [CB12, GLL*10]). A few other notable tools can be found that use visualization to a greater extent, such as those by Chuang et al. [CMH12] and Gretarsson et al. [GOB*12]. Still, much room is left for further exploration of how visualization techniques can improve communication of topic modeling results, and results from text mining models in general. This exploration is also supported by Tufte's [Tuf83] design principle, to use graphics when words alone cannot communicate the message effectively, as structural information is as central to the analysis of text as semantics.

Our work extends that of Chuang et al. [CMH12], which we find to be the most promising previous example of topic model visualization. They visualize topic-keyterm relations through a matrix view, which provides some idea of topic distribution, similarities and meaning. Our approach may be seen as a graph visualization that uses a topic-keyterm matrix similar to theirs as an adjacency matrix. We argue that the force-directed graph visualization provides a view that more intuitively communicates topic similarity structure. Compared to the visualization of Liu et al. [LZP*09], we aim to show topic relations more explicitly and detailed.

3. Visualization of Topic Models using Graphs

The raw topic-term probabilities provided by LDA are not suitable as a ranking to find terms that distinguish well between topics. Various measures have been proposed for reranking that provide better description of the topic meaning, such as a TFIDF-inspired *term-score* [BL09] and others that likewise penalize terms that are prevalent in many topics. Reranking the terms is essential to produce descriptive and distinguishing keyterms for presenting topic meaning.

For the sake of interpretability, we use the conditional probability $P(T|w)$ to score how distinguishing a term is of its topic. Given a term w is observed, the probability of it belonging to topic T is a measure of how distinguishing w is of T. We derive the measure as $P(T|w) = P(w|T)P(T)/P(w)$ where $P(w|T)$ is the topic-term probability distribution of LDA. Using $P(T|w)$ is in line with [CMH12]. A threshold on the probability selects the top keyterms for each topic.

As our basic visualization technique, we use a graph with force-directed layout (using the Barnes-Hut algorithm [BH86]). It provides a spatial metaphor for topic similarity in the corpus. The graph consists of topic nodes connected to keyterm nodes only, which produces a fairly sparse graph for which force-directed layouting can produce clear results (confer online demo). Topics are not directly connected to each other, rather only through their common keyterms. Still, the node positioning communicates general topic similarity structures, and the connecting keyterms provide qualitative detail on the nature of their relation. In the topic overview, keyterm weighting is encoded by link opacity to communicate the strength of the relation, node size is relative to the general frequency of the topic and colors from a qualitative scale are used to better distinguish the topic neighborhoods.

Nodes can be dragged to alter their positions, which are then adjusted by the force-directed algorithm in real time. However, the automatic adjustment is slow enough to allow the user to move several nodes before a stable conformation is reached, which allows for interactive exploration of alternative locally optimal conformations and helps the user inspect the structure of the graph. Zooming and panning supports exploration of many topics even on smaller screens.

The meaning of a topic is represented by the weighted keyterms linked to it. While the initial view provides an overview of all topics, focusing on a single topic by hovering highlights its details. Highlighting fades all parts of the graph not connected to the topic to preattentively direct the user and ease their inspection. It provides context plus focus [CMS99] and follows the visual information seeking mantra [Shn96] by filtering information and providing details on demand. The topic-specific term weighting can now also be encoded through keyterm font size, which enables simultaneous reading of term meaning and importance. Similar to a tag cloud, this creates an easy-to-read *topic cloud*.

Some terms are shared among topics, hinting at their ambiguity. Ideally, a topic represents a meaningful context in which a term is used, through which the sense of the term is clearified. Two topics sharing a term might indicate that the term holds two different meanings in the corpus. Handling such ambiguity is a central purpose of topic models, but simple representations do not facilitate identification of such patterns by the user, while visualization easily can.

As the graph displays only topic nodes and a limited set of keyterm nodes, but no document information, it acts as an abstracted topical view of the corpus. All document information is accessible only through interaction, which means that scaling the corpus size does not affect the visualization. However, information retrieval functionality is not discussed here due to space constraints.

References

[BH86] BARNES J., HUT P.: A hierarchical o (n log n) force-calculation algorithm. *Nature 324* (1986), 446–449. 2

[BL09] BLEI D. M., LAFFERTY J. D.: Topic models. In *Text mining: Classification, clustering, and applications*. Chapman & Hall/CRC Press, Boca Raton, FL, 2009, ch. 10, p. 71. 1, 2

[Ble12] BLEI D. M.: Probabilistic topic models. *Communications of the ACM 55*, 4 (2012), 77–84. 1

[CB12] CHANEY A. J.-B., BLEI D. M.: Visualizing topic models. In *ICWSM* (2012). 2

[CMH12] CHUANG J., MANNING C. D., HEER J.: Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (2012), ACM, pp. 74–77. 2

[CMS99] CARD S. K., MACKINLAY J. D., SHNEIDERMAN B.: *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999. 2

[GLL*10] GARDNER M. J., LUTES J., LUND J., HANSEN J., WALKER D., RINGGER E., SEPPI K.: The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization* (2010). 2

[GOB*12] GRETARSSON B., O'DONOVAN J., BOSTANDJIEV S., HÖLLERER T., ASUNCION A., NEWMAN D., SMYTH P.: Topicnets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST) 3*, 2 (2012), 23. 2

[LZP*09] LIU S., ZHOU M. X., PAN S., QIAN W., CAI W., LIAN X.: Interactive, topic-based visual text summarization and analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), ACM, pp. 543–552. 2

[RKPW08] RISCH J., KAO A., POTEET S. R., WU Y.-J. J.: Text visualization for visual text analytics. In *Visual data mining*. Springer, 2008, pp. 154–171. 1

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (Boulder, CO, 1996), pp. 336–343. 2

[Tuf83] TUFTE E.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983. 2

[War04] WARE C.: *Information Visualization: Perception for Design*. Morgan Kaufman, San Francisco, CA, 2004. 1

Paper IV

IV

Exploratory topic modeling with distributional semantics

S. Rönnqvist (2015). In *Advances in Intelligent Data Analysis XIV*, Lecture Notes in Computer Science 9385, pages 241–252. Springer

Exploratory Topic Modeling with Distributional Semantics

Samuel Rönnqvist(✉)

Turku Centre for Computer Science – TUCS,
Department of Information Technologies, Åbo Akademi University, Turku, Finland
sronnqvi@abo.fi

Abstract. As we continue to collect and store textual data in a multitude of domains, we are regularly confronted with material whose largely unknown thematic structure we want to uncover. With unsupervised, exploratory analysis, no prior knowledge about the content is required and highly open-ended tasks can be supported. In the past few years, probabilistic topic modeling has emerged as a popular approach to this problem. Nevertheless, the representation of the latent topics as aggregations of semi-coherent terms limits their interpretability and level of detail.

This paper presents an alternative approach to topic modeling that maps topics as a network for exploration, based on distributional semantics using learned word vectors. From the granular level of terms and their semantic similarity relations global topic structures emerge as clustered regions and gradients of concepts. Moreover, the paper discusses the visual interactive representation of the topic map, which plays an important role in supporting its exploration.

(Topic mapping code and demo available at http://samuel.ronnqvist.fi/topicMap/)

Keywords: Topic modeling · Distributional semantics · Visual analytics

1 Introduction

Following the increase in digitally stored and streamed text, the interest for computational tools that aid in organizing and understanding written content at a large scale has soared. Natural language processing and machine learning techniques demonstrate strength in their feats of handling the challenging intricacies of human language to extract information and in their aptitude for scanning big data sets. However, while we can model what information is likely to be interesting, humans alone are capable of a deeper understanding that involves evaluating information against a wide and diverse body of knowledge in nuanced ways, which motivates a focus on human-computer interaction and visual analytics in text mining [17].

This paper concerns analysis of text by means of *exploratory topic modeling*, by which I emphasize the exploratory use of models that convey topic structure.

E. Fromont et al. (Eds.): IDA 2015, LNCS 9385, pp. 241–252, 2015.
DOI: 10.1007/978-3-319-24465-5_21

To this end, I put forward a new method for topic modeling based on distributional semantics using continuous word vector representations, for the construction of models called *topic maps*. On the one hand, the focus is set explicitly on unsupervised learning to allow maximum coverage in terms of domain and language without need for adaption, while taking advantage of recent advances in word vector training by neural networks. On the other hand, the role of the human user is acknowledged as an important part of the analysis process as the one who understands and explores the modeling results; therefore, visual interactive presentation is discussed as part of the contribution alongside map construction and perceived as equally important to exploratory topic modeling.

Probabilistic topic modeling [4] is a family of machine learning algorithms for uncovering thematic structure in text documents that are widely used, and applicable both for exploratory analysis of topics and as a discrete dimensionality reduction method in support of other learning tasks. Based on co-occurrence of terms in documents, probabilistic topic modeling extracts a number of latent topics. In the seminal algorithm, Latent Dirichlet Allocation (LDA), the number of topics to infer is given as a parameter. Assuming that each document may discuss a mixture of topics, it attempts to isolate coherent topics. Each topic is defined as a probability distribution over terms, where the terms collectively carry the meaning of the latent topic. While LDA and many of its variations are theoretically solid and rest on an interpretation-friendly probabilistic basis, issues of interpretability are nevertheless commonplace and well recognized [6]. First, the unsupervised modeling offers no guarantees that the topic division is semantically meaningful; some topics may seem similar and hard to distinguish, whereas others turn out very specific. These issues may be mitigated by selecting appropriate parameters, including the number of topics in the case of LDA. Second, the terms within topics may appear semantically incoherent and confusing to a human. Various efforts have been made to improve coherence (e.g., [14,15]), yet for humans to form an understanding of what a topic signifies based on a set of weighted terms, interpretation inevitably involves a certain cognitive load, only increased by the iterative task of contrasting topics against each other to grasp the broader picture.

Thoughtful visual representation of the topic structure and terms can ease the task (see, e.g., [7,18]), but I argue that in many cases it is more meaningful to choose to operate from the level of individual terms that represent concrete concepts and their bilateral semantic similarity relations. A discrete division of topics is practical in many use cases, but is somewhat unnatural for exploratory purposes, and mere aggregation of terms inevitably leads toward less interpretable abstractions. Instead, it is more fitting to allow for a topic structure to emerge as a global property from the local semantic similarity relations among terms. Such a semantic network allows the human user to flexibly identify topics as regions through proper visualization, while the network also supports quantitative analysis such as community detection [9] (overlapping clustering, which handles ambiguous terms) to identify discrete topics.

The following section introduces the method for building the semantic network model, the topic map, whereas Sect. 3 discusses its visualization, and Sect. 4 reports on experiments conducted to demonstrate the mapping method, followed by some concluding remarks.

2 Building the Topic Map

Distributional semantics models the meanings of words based on their contexts, namely the surrounding words in a sentence, according to the aphorism "you shall know a word by the company it keeps" [8]. While modeling has traditionally been based on counting of context words, recent approaches that work by learning to predict words instead have been highly successful [1]. A popular way of representing the semantics is by vectors, e.g., through projection [19] or later through neural network training [3]. Lately, Mikolov et al. [13] have shown how neural networks can be efficiently used to train semantic models based on corpora at the scale of billions of words, in order to achieve very high semantic accuracy. Their continuous skip-gram model is a neural network trained to predict context words based on the center word, using a single hidden layer. Through supervised training, the network optimizes its hidden layer weights, which results in the learned array of hidden nodes providing fixed-length vector representations of word semantics, i.e., word vectors. The word vectors embed words into a semantic space that supports measuring similarities among words by their vectors (e.g., by cosine similarity), as well as other vector arithmetic operations (e.g., addition and subtraction for regularities prediction).

For the purpose of modeling the general topic composition of corpora, I use the neural network skip-gram method to model word-level semantic similarity, and from pairwise relations let the broader topic structure emerge. Whereas the focus in word vector training generally is to approximate the semantics of language in general, which can be achieved by training on large and diverse enough text, the idea is here to explicitly model the semantics of the language in one's corpus alone. The model then reflects how words relate in the discourse of the corpus rather than elsewhere. Thereby, the discrepancies between the word similarities presented by the model and the observers own, more general understanding and less data-informed expectation of how the words relate, constitute telltales of the thematic nature of the underlying text. (Kulkarni et al. use word vectors accordingly to study linguistic change in English over time [11].) For topic modeling to be meaningful, it naturally needs to work for corpora far smaller than billions of words. As will be demonstrated in Sect. 4, skip-gram models can learn usefully accurate word vectors on much smaller data sets, too.

Apart from semantic similarity, the topic map incorporates term frequencies, used to represent the prevalence of terms in the corpus, and in combination with their semantic neighborhood provide a sense of the overall importance of sections of the map, reflecting the prevalence of specific concepts or topics. Probabilistic topic modeling similarly uses topic-wise probability distributions over terms to represent their degree of importance within the topic.

Algorithm 1. Topic map construction (in: tokens, V, C, E, N, P, L; out: net)

```
# Word vector training
model = Word2Vec(tokens, vector_size=V, context_size=C, epochs=E)
# Network construction
for i1 in range(0, N-1):
  for i2 in range(i1+1, N):
    t1, t2 = top_N_terms[i1], top_N_terms[i2]
    net.add_link(t1, t2, weight=model.similarity(t1, t2))
# Network pruning
threshold = percentile([link.weight for link in net.links], P)
for node in net.nodes:
  cap = sorted(net.links[node], key=lambda link: link.weight)[-1*L].weight
  for link in net.links[node]:
    if link.weight < max(cap, threshold):
      net.remove_link(link)
ws = [link.weight for link in net.links]
for link in net.links:
  net.links[link].weight = (link.weight-min(ws)) / (max(ws)-min(ws))
```

Using the word vector model and term counts, a semantic network that constitutes the topic map can be constructed according to Algorithm 1 as described in the following. First, the text of a corpus is processed and tokenized into meaningful and well normalized terms. Then, the map is constructed through the following two main steps.

Word Vector Training. Given the main parameters, vector size (V) and context size (C), word vectors are trained on term sequences by the method of Mikolov et al. (word2vec). Vector size determines the dimensionality of the semantic space and is customarily in the range of 50 to 1000, where higher dimensionality allows for a finer model given enough data. The size of the word context to consider is typically about 5–10 words, but for the current task even contexts up to 25 words have proved satisfactory. Training in multiple epochs (E) (e.g., 3–10) tends to improve the quality of the model noticeably, especially with little data available.

Network Construction. Once the vectors have been trained, we can use the model to measure similarity of pairs of terms. The most frequent terms in the corpus are picked for comparison, preferably excluding stopwords. Typically in the range 100–1000, the number of unique terms to include (N) defines the maximum level of detail in the topic map and limits the computational complexity of building it. For each pair, the cosine similarity between their vectors ($sim(t_1, t_2) = \boldsymbol{v}(t_1) \cdot \boldsymbol{v}(t_2)$, with unit vectors) is computed and stored.

Network Pruning. As only similar terms are meaningful to relate and as we seek to build a network that is neither too dense and cluttered nor too sparse and disconnected, the pairs with highest similarity scores are retained as links between the term nodes. With varying sizes of the vector and corpora, the similarity scores vary considerably as well. Thus, filtering of pairs is performed by

a threshold defined as a percentile of all scores stored (P), typically at the 97–99$^{\text{th}}$ percentile, which makes the parameter's effect more stable. Moreover, an upper bound on number of links per term (L) helps reduce cluttering density due to general terms that may measure as very similar to many terms. Typical cap values are 8–15 links per term. All links are finally weighted according to its normalized similarity score, as a standard measure of link strength ($w' \in [0, 1]$).

In order to optimize parameter selection the quality of the topic maps must be evaluated. While the exploratory task ultimately calls for qualitative evaluation, semantic prediction accuracy will be used for initial guidance in word vector training, which is the more computationally demanding step. The evaluation method and data, borrowed from Mikolov et al., measures syntactic and semantic regularities such as "man is to woman as king is to *queen*", "Athens is to Greece as Baghdad is to *Iraq*" and "code is to coding as dance is to *dancing*", where accuracy in predicting the last word is evaluated. Measuring how well the model approximates general English, the relative performance on this task can help to rule out models that are too simple and produce suboptimal maps because they lack ability to appropriately model the semantics. The highest accuracy, however, does not necessarily provide the best topic map, as its quality relies on a balance between specificity and generality of its relations. The experiments in Sect. 4 illustrate this further.

Apart from local link accuracy, the network should ideally show good structure in terms of how broader clusters emerge, too. This is highly dependent on both calibration of the network parameters and how the network is analyzed. The experiments in this paper focus on visual analysis based on force-directed layouting, in which case desirable network structures contain some degree of clustering into coherent and meaningful regions, without excessive cross-linking between terms in different clusters to avoid overlaps. The network construction parameters (P, L, N) may be adjusted to optimize the readability of the map, which in practice can be done instantaneously while visualizing the network. Hence, optimization of the word vector parameters is the more cumbersome groundwork that begets good maps, and evaluation of accuracy helps by reducing the search space.

3 Visualizing the Topic Map

Exploration of complex models such as topic models calls for presentations that provide as much detail as meaningfully possible. The most information-dense mode of communication is visualization, whereas interactivity helps expand the space of information that can be presented intelligibly on a finite screen. The visual analytics paradigm [10] embraces visual interactive interfaces as they offer a means of communication that is both rich and reactive, thus, helping users in making sense of models and data. Visualization of the topic map incorporates Shneiderman's visual information-seeking mantra, "overview first, zoom and filter, then details-on-demand" [20], by providing both overview of a corpus and a scaffold for exploration of its details. Visualization of the two main aspects of

the map, term frequencies and word vectors, is discussed in the following, as well as their combination into a visual topic map.

Among the most popular forms of text visualization are word clouds, which are simple yet useful. Their main property, representing word importance by size, is powerful because it utilizes a preattentively recognized visual variable, i.e., relative word importance is recognized effortlessly and without requiring focused attention, in parallel across the field of vision at the early stage of the human visual system [21]. While word clouds have received some criticism relating to other properties such as the (dis)organization of words, studies have sought improvement and in terms of readability evaluated various approaches such as clustered [12] and semantic word clouds [2] that impose some semantically meaningful organization of words. However, so far none of the approaches have used distributional semantics, which offers advantages by being arguably more specific than tried clustering approaches, and unsupervised as opposed to database approaches.

By contrast, a common approach to visualizing word vectors is to plot words according to their two-dimensional projections by PCA (or other multidimensional scaling methods), which achieves a basic form of semantic organization, albeit easily cluttered at the center. While word clouds commonly place words as closely as possible, regarding order or not, projection uses planar distance to communicate the degree of semantic similarity (a spatial visual metaphor) as well as ordering. Nevertheless, projection into two dimensions is bound to produce overlap between semantically unrelated sets of words, which motivates the visualization of semantic relations by drawn line connections that is more explicit [16]. For visualization of the topic map network, I propose to use a force-directed layout (a projection method) that optimizes word positions explicitly based on semantic relations present in the network model, rather than the whole word vector model. In particular, the D3 force algorithm [5] is suitable as it can counter overlap of terms (by node charge) to preserve readability, even when they are densely connected. It can also run in real time to allow for interactive adjustment of positions which lets the user explore multiple local optima of positioning.

The visual topic map lends from word clouds the word sizing relative to their corpus frequency, and uses force-directed layouting to organize the map semantically. Drawing the network of words, the strength of each link is encoded by opacity, which makes more explicit the relative importance of individual links, and together with the emergent density of links it provides an aggregate impression of the varying density of the map.

Interactive exploration of the map is enabled foremost by zoom/pan capabilities, which in a very direct way allows more terms to be displayed, and highlighting of links of specific terms. The filtering of terms by frequency can be responsive to the level of zoom to seamlessly provide more detail on demand. The percentile filter used to construct the network can be relaxed if the visual interface can counter the added complexity, and the number of terms can be increased accordingly. Hence, the scalability of the topic map visualization depends largely

on interaction design. The semantic network of frequent terms also functions as a canvas for other types of information, such as mapping of local term neighborhoods and relational information, touched upon in Sect. 5.

4 Experiments

The topic map will be demonstrated and tested using two different corpora. The first corpus is a sample of news articles from Reuters (U.S. online edition) containing 23 k articles and 9.2 M words (167 k unique) and the other is a collection of financial patent application abstracts from the U.S. Patent and Trademark Office comprising 14 k abstracts and 1.7 M words (20 k unique). In accordance with the exploratory aim of this article, a topic map is trained and visualized for each corpus as discussed above, in hope of gaining insight into the thematic composition of each (see Figs. 1 and 2). The most prevalent terms representing concrete concepts are displayed and their semantic similarity relations provide organization that portray topics implicitly as regions and gradients between them.

The experiment starts by selecting the parameters for word vector training, guided by quantitative accuracy and qualitative assessment of the map (as discussed in Sect. 2). Having exhaustively tested various settings, their relationship

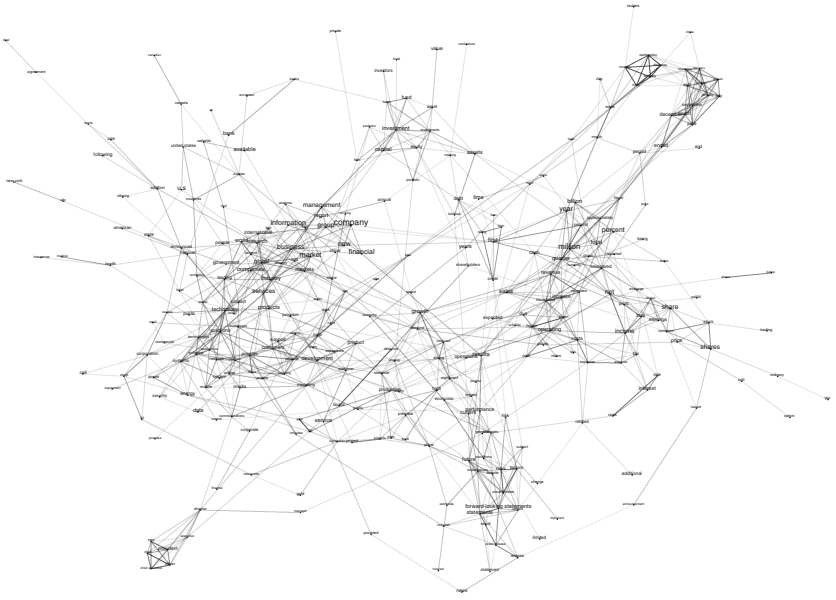


Fig. 1. Topic map of financial news articles, interactive version available at: http://samuel.ronnqvist.fi/topicMap/

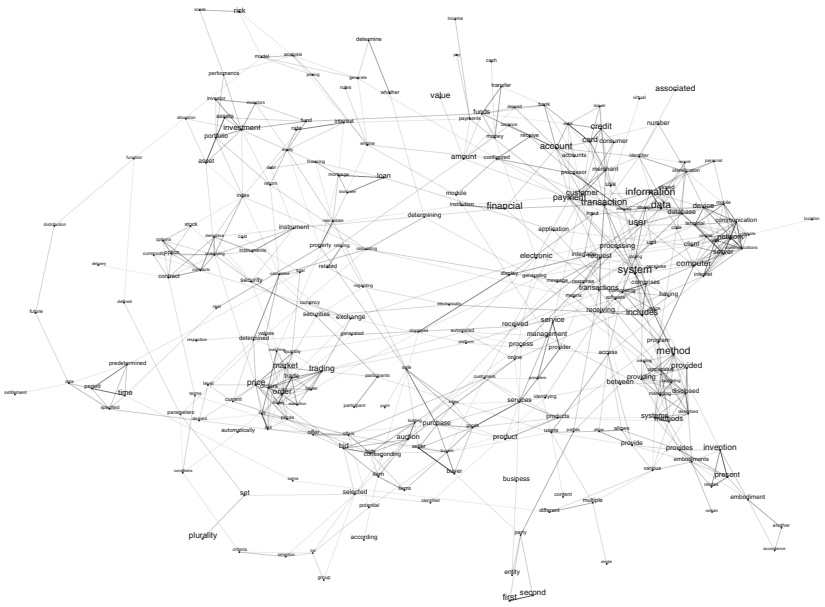


Fig. 2. Topic map of financial patent abstracts

can be described as follows. With a fixed context size of 15 for the Reuters corpus, accuracy reaches a plateau from vector sizes 200 to 500 (on average at 17 %, with $E = 3$), decreasing afterwards. Meanwhile, at a given vector size, accuracy tends to asymptotically approach a limit with increased context size. Qualitatively, the best network structure appears to result from settings where accuracy is close to the limit but context size is kept moderate.

The experiments show that the map is surprisingly robust with respect to the training parameters, producing largely comprehensible results even at vector sizes of 25 or 600 and context sizes of 5 and 50 respectively. Nevertheless, the quality of the Reuters map is noticeably best at vector sizes 200–400 and context sizes 10–20, where larger contexts benefit from larger vectors. Simpler models produce networks with smaller regions that are tightly clustered, but result in either few or arbitrary connections between regions, depending on the threshold (P). Networks from complex models have similar problems, although the strong connections tend to be very specific and semantically accurate, which explains their good testing performance.

The qualitatively optimal models in between strike a balance between, on the one hand, semantic accuracy that provides a map of meaningful connections and, on the other hand, generality by connecting parts of the map through more abstract but still helpfull term relations. Hence, measured accuracy provides fundamental guidance in learning a model that handles the language well, but the map then benefits from a slight regularizing or smoothing effect achieved by

using a simpler model than the quantitatively optimal. While large vectors and contexts combined can achieve maximum accuracy (about 22 % for the Reuters corpus), it does not seem productive to surpass contexts of about 25 words, and given a limited context size, it is motivated to choose a vector size towards the beginning of the accuracy plateau. The number of training epochs has a strong effect on accuracy, e.g., the settings $V = 400$, $C = 15$ and $E = \{1, 3, 5\}$ give accuracies 7.3, 16.7 and 19.1, but the two latter cases do not show any notable qualitative difference for the Reuters data.

The topic map in Fig. 1 was produced with the settings $V = 250$, $C = 12$, $E = 5$, $N = 500$, $P = .985$ and $L = 12$ (accuracy 17.6 %, training time 14.5 min on 4 cores). It depicts the topic landscape of the Reuters financial news corpus by its most frequent terms excluding stop words (including automatically detected bi-gram phrases). The similarity threshold set at the 98.5^{th} percentile provides an appropriate degree of connectivity. The cap on links per term helps improve readability especially in the dense region surrounding the terms *business* and *technology*. The map uncovers an uneven distribution of terms, where smaller concentrations highlight cliques of terms (e.g., president, ceo, etc. down left) that represent a rather distinct general concept. Larger concentrated regions form to highlight a broader topic division of the corpus, the three main regions broadly reflecting discourse on business-related activities, realized performance and expected performance.

The map in Fig. 2 similarly illustrates the lay of specific concepts and more general topics as they occur in the set of patent abstracts. A few themes can be identified, such as payment systems, telecommunications, trading, portfolio management and patent-specific language. The map includes 350 terms and links for the top 2 % most similar pairs. As the patent corpus is much smaller the vector size was reduced according to vocabulary size heuristically by $\frac{V_1^2}{|vocab_1|} \approx \frac{V_2^2}{|vocab_2|}$ to $V = 85$, context size was kept at 12 not to reduce the already scarce data and training was run in 10 epochs (training time 3.2 min on 4 cores).

To conclude the evaluation of the generated topic maps, I compare the news corpus against a benchmark obtained by LDA (results for the patent corpus are similar but omitted due to space constraints). The same preprocessing of the text is used as above, and the topics are modeled with standard parameter settings into 8 topics. Each topic is presented by their top-10 terms according to the topic-term probability distributions, as the most direct way of presenting the model. Stop words are excluded to make the results more informative. While several methods have been proposed that rerank terms to better support interpretation of the topics (cf. [7,18,22]), no such method seems to have been unanimously or widely adopted. The obtained topics are:

Topic 0: million, net, quarter, year, financial, income, company, share, operating, total
Topic 1: securities, class, relevant, number, options, option, price, form, code, relevant security
Topic 2: company, shares, fitch, fund, rating, share, ratings, information, financial, available
Topic 3: u.s, bank, new, company, financial, government, state, group, year, years
Topic 4: first, people, world, new, patients, home, years, health, year, games
Topic 5: company, information, new, services, business, market, products, forward-looking statements, technology, solutions
Topic 6: q2 2014, jul amc, call, company, 29 jul, corp, earnings conf, jul bmo, trust, share
Topic 7: percent, year, million, billion, market, u.s, sales, shares, growth, down

For some topics it is possible to discern a latent meaning, while others prove hard to interpret. For instance, Topics 0 and 7 appear to relate to realized financial performance, but it is difficult both to form a more detailed explanation of them and to distinguish logically between them. As mentioned in Sect. 1, recognizing a distinct topic from an aggregate of terms is challenging, as is the task of understanding how multiple topics relate. While the topic map includes many of the same frequent terms, its natural, semantic organization makes it easier to view and grasp the overall topic composition and scope. Local neighborhoods of the map tend to be more coherent than LDA topics, and the relation between different sections of the map is made more explicit. While exploration of LDA topic models can be supported by meaningful presentation (e.g., [7,18]), the topic map's alternative way of approaching topic modeling remains well motivated for exploration.

5 Discussion

My aim has been to introduce a new approach of using distributional semantics, specifically word vectors trained by neural networks, to explore topics in bodies of text. A problem commonly addressed by probabilistic topic modeling, this approach sets out to tackle it with finer granularity, by building a topic map bottom-up from concrete terms towards general topics, rather than forcing interpretation of implicit meaning among an explicit, but not necessarily coherent, set of topic terms. Distributional-semantic modeling provides meaningful word-to-word similarity relations and organization that is easy to navigate. In addition, I put forward a visualization design for the map that provides overview and means for linking to further details, thus supporting interactive exploration. As a network model, the map also supports quantitative network analysis, in particular community detection as a form of second-level clustering to provide explicit topics, which are useful in some cases. The topic map opens up to a range of possible extensions to be explored.

As the map provides a projection of the semantic space of a corpus, another interesting type of information is the relational, i.e., how different concepts are referenced together in text. Mapping such relations onto the topic map may lead to still more informative ways of summarizing the contents of texts. Document-level co-occurrence of terms used in probabilistic topic modeling represents a crude way of harnessing relational information to extract topic information, but it is likely beneficial to treat distributional word context similarity and word-to-word co-occurrence as separate aspects that both contribute toward summarizing the discourse of a corpus. Thus, the approach of constructing a topic map outlined in this paper should be seen as elementary to future extensions that among other things include sophisticated analysis of relations in text and powerful visual interactive interfaces to make the semantic space and its linked information readily browsable. The semantic network is the basic data structure, which can be meaningfully presented in many other ways as well, e.g., using more structured network layouts or non-graphical representation, possibly emphasizing search with a completely local focus rather than overview.

Studying immediate neighborhoods of specific terms may in fact be a desirable mode of exploration, which can be supported in other ways than described above. Rather than starting from the frequent term set, terms with the closest vectors can be searched. By recursively traversing the nearest neighbors of a term, a close-up view of its semantic context in the corpus is obtainable.

Vector similarity comparisons can also be performed with compound vectors that average two or a few word vectors, for instance, as a way to disambiguate a term (e.g.: *financial* by *financial+group*, *financial+results*) or merge closely related terms (e.g., *customer+customers*). The latter could be applied to enhance the map by reducing term redundancy and thereby visual clutter, while joining their term counts. Another way to generalize across terms would be to smooth term counts to some extent among direct neighbors, in order to make the representation of prevalence of regions more congruent.

In this paper, word vectors and term frequencies were obtained from the same set of text, which may lead to problems of accuracy for the study of smaller sets of text (e.g., in the order of 10–100 k rather than 1 M words). It is possible to separate these, letting the word vectors be trained on a larger background corpus while counting terms on a smaller foreground set, as long as they are related in nature. For instance, the background corpus may consist of text from a single source over a certain period of time, while texts from smaller intervals during that period would be used as foreground corpora to allow for more specific study of varying term prevalence over time, still benefiting from a more robust semantic model.

As efficient word vector training with neural networks has opened up many new possibilities in natural language processing, I hope to introduce it for the purpose of exploring topics in masses of text by proposing a methodology for building and visualizing topic maps. Unsupervised word-level modeling of semantics offers very flexible and detailed means for analysis that deserve further study. The concluding discussion has outlined a few interesting future directions, and ultimately the utility of topic maps and their visual representations should be tested by how they support users' understanding in a variety of real-world settings.

References

1. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 238–247 (2014)
2. Barth, L., Kobourov, S.G., Pupyrev, S.: Experimental comparison of semantic word clouds. In: Gudmundsson, J., Katajainen, J. (eds.) SEA 2014. LNCS, vol. 8504, pp. 247–258. Springer, Heidelberg (2014)
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
4. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
5. Bostock, M., Ogievetsky, V., Heer, J.: D3: data-driven documents. IEEE Trans. Vis. Comp. Graph. **17**(12), 2301–2309 (2011). (Proc. InfoVis)

6. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)
7. Chuang, J., Manning, C.D., Heer, J.: Termite: visualization techniques for assessing textual topic models. In: Advanced Visual Interfaces (2012)
8. Firth, J.: A synopsis of linguistic theory 1930–1955. In: Studies in Linguistic Analysis, pp. 1–32. Philological Society, Oxford (1968)
9. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
10. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: scope and challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
11. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. arXiv preprint arXiv:1411.3315 (2014)
12. Lohmann, S., Ziegler, J., Tetzlaff, L.: Comparison of tag cloud layouts: task-related performance and visual exploration. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 392–404. Springer, Heidelberg (2009)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at International Conference on Learning Representations (2013)
14. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, pp. 262–272 (2011)
15. Newman, D., Bonilla, E.V., Buntine, W.: Improving topic coherence with regularized topic models. In: Advances in Neural Information Processing Systems 24, pp. 496–504 (2011)
16. Palmer, S., Rock, I.: Rethinking perceptual organization: the role of uniform connectedness. Psychon. Bull. Rev. **1**(1), 29–55 (1994)
17. Risch, J., Kao, A., Poteet, S.R., Wu, Y.-J.J.: Text visualization for visual text analytics. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 154–171. Springer, Heidelberg (2008)
18. Rönnqvist, S., Wang, X., Sarlin, P.: Interactive visual exploration of topic models using graphs. In: Eurographics Conference on Visualization (EuroVis) (2014)
19. Schütze, H.: Dimensions of meaning. In: Proceedings of the 1992 ACM/IEEE Conference on Supercomputing, Supercomputing 1992, pp. 787–796 (1992)
20. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the IEEE Symposium on Visual Languages, pp. 336–343 (1996)
21. Treisman, A.: Features and objects in visual processing. Sci. Am. **255**(5), 114–125 (1986)
22. Wilson, A.T., Chew, P.A.: Term weighting schemes for latent dirichlet allocation. In: Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 465–473 (2010)

Paper V

V

Do we really need all those rich linguistic features? A neural network-based approach to implicit sense labeling

N. Schenk and C. Chiarcos and K. Donandt and S. Rönnqvist and E.A. Stepanov and G. Riccardi (2016). In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL)*, pages 41–49. Association for Computational Linguistics

Do We Really Need All Those Rich Linguistic Features? A Neural Network-Based Approach to Implicit Sense Labeling

Niko Schenk[*], **Christian Chiarcos**[*], **Kathrin Donandt**[*],
Samuel Rönnqvist[*,†], **Evgeny A. Stepanov**[‡] and **Giuseppe Riccardi**[‡]
[*]Applied Computational Linguistics Lab, Goethe University, Frankfurt am Main, Germany
[†]Turku Centre for Computer Science, TUCS, Åbo Akademi University, Turku, Finland
[‡]Signals and Interactive Systems Lab, DISI, University of Trento, Italy
{schenk,chiarcos,donandt}@informatik.uni-frankfurt.de,
sronnqvi@abo.fi, {evgeny.stepanov,giuseppe.riccardi}@unitn.it

Abstract

We describe our contribution to the CoNLL 2016 Shared Task on shallow discourse parsing.[1] Our system extends the two best parsers from previous year's competition by integration of a novel *implicit* sense labeling component. It is grounded on a highly generic, language-independent feedforward neural network architecture incorporating weighted word embeddings for argument spans which obviates the need for (traditional) hand-crafted features. Despite its simplicity, our system overall outperforms all results from 2015 on 5 out of 6 evaluation sets for English and achieves an absolute improvement in F_1-score of 3.2% on the PDTB test section for non-explicit sense classification.

1 Introduction

Text comprehension is an essential part of Natural Language Understanding and requires capabilities beyond capturing the lexical semantics of individual words or phrases. In order to understand how meaning is established, altered and transferred across words and sentences, a model is needed to account for contextual information as a semantically coherent representation of the logical *discourse structure* of a text. Different formalisms and frameworks have been proposed to realize this assumption (Mann and Thompson, 1988; Lascarides and Asher, 1993; Webber, 2004).

In a more applied NLP context, *shallow discourse parsing* (SDP) aims at automatically detecting relevant discourse units and to label the relations that hold between them. Unlike *deep discourse parsing*, a stringent logical formalization or the establishment of a global data structure, for instance, a tree, is not required.

With the release of the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB), annotated training data for SDP has become available and, as a consequence, the field has considerably attracted researchers from the NLP and IR community. Informally, the PDTB annotation scheme describes a discourse unit as a syntactically motivated character span in the text, augmented with relations pointing from the second argument (*Arg2*, prototypically, a discourse unit associated with an explicit discourse marker) to its antecedent, i.e., the discourse unit *Arg1*. Relations are labeled with a relation type (its *sense*) and the associated discourse marker (either as found in the text or as inferred by the annotator). PDTB distinguishes *explicit* and *implicit* relations depending on whether such a connector or cue phrase (e.g., *because*) is present, or not.[2] As an illustrative example without such a marker, consider the following two adjacent sentences from the PDTB:

Arg1: *The real culprits are computer makers such as IBM that have jumped the gun to unveil 486-based products.*
Arg2: *The reason this is getting so much visibility is that some started shipping and announced early availability.*

In this *implicit* relation, *Arg1* and *Arg2* are directly related. The discourse relation type is *Expansion.Restatement*—one out of roughly twenty finegrained tags marking the sense relation

[1] http://www.cs.brandeis.edu/~clp/conll16st
Our parser code is available at: https://github.com/acoli-repo/shallow-discourse-parser

[2] The set of relation types is completed by alternative lexicalization (*AltLex*, discourse marker rephrased), entity relation (*EntRel*, i.e., anaphoric coherence), resp. the absence of any relation (*NoRel*).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 41–49,
Berlin, Germany, August 7-12, 2016.

between any given argument pair in the PDTB.

Our Contribution: We participate in the CoNLL 2016 Shared Task on SDP (Xue et al., 2016; Potthast et al., 2014) and propose a novel, neural network-based approach for implicit sense labeling. Its system architecture is modular, highly generic and mostly language-independent, by leveraging the full power of pre-trained word embeddings for the SDP sense classification task. Our parser performs well on both English and Chinese data and is highly competitive with the state-of-the-art, though does not require manual feature engineering as employed in most prior works on implicit SDP, but rather relies extensively on features learned from data.

2 Related Work

Most of the literature on automated discourse parsing has focused on specialized subtasks such as:

1. **Argument identification**
   (Ghosh et al., 2012; Kong et al., 2014)
2. **Explicit sense classification**
   (Pitler and Nenkova, 2009)
3. **Implicit sense classification**
   (Marcu and Echihabi, 2002; Pitler et al., 2009; Lin et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014)

A minimal requirement for any full-fledged end-to-end discourse parser is to integrate at least these three processes into a sequential pipeline. However, until recently, only a handful of such parsers have existed (Lin et al., 2014; Biran and McKeown, 2015; duVerle and Prendinger, 2009; Feng and Hirst, 2012). It has been enormously difficult to evaluate the performance of these systems among themselves, and also to compare the efficiency of their individual components with other competing methods, as i.) those systems rely on different theories of discourse, e.g., PDTB or RST; and ii) different (sub)modules involve custom settings, feature- and tool-specific parameters, (esp. for the most challenging task of *implicit sense labeling*). Furthermore, iii) most previous works are not directly comparable in terms of overall accuracies as their underlying evaluation data suffers from inconsistent label sizes among studies (e.g., full sense inventory vs. simplified 1- or 2-level classes, cf. Huang and Chen (2011)).

Fortunately, with the first edition of the shared task on SDP, Xue et al. (2015) had established a *unified framework* and had made an independent evaluation possible. The best performing participating systems – most notably those by Wang and Lan (2015) and Stepanov et al. (2015) – have re-implemented the well-established techniques, for example the one by Lin et al. (2014).

2.1 Deep Learning Approaches to SDP

In last year's shared task, first implementations on *deep learning* have seen a surge of interest: Wang et al. (2015) and Okita et al. (2015) proposed a recurrent neural network for argument identification and a paragraph vector model for sense classification. Distributed representations for both arguments were obtained by vector concatenation of embeddings.

An earlier attempt in a similar direction of *representation learning* (Bengio et al., 2013) has been made by Ji and Eisenstein (2014). The authors demonstrated successfully how to discriminatively learn a latent, low-dimensional feature representation for RST-style discourse parsing, which has the benefit of capturing the underlying meaning of elementary discourse units without suffering from data sparsity of the originally high dimensional input data.

Closely related, Li et al. (2014) introduced a recursive neural network for discourse parsing which jointly models distributed representations for sentences based on words and syntactic information. The approach is motivated by Socher et al. (2013) and models the discourse unit's root embedding to represent the whole discourse unit which is being obtained from its parts by an iterative process. Their system is made up of a binary structure classifier and a multi-class relation classifier and achieves similar performance compared to Ji and Eisenstein (2014).

Very recently, Liu et al. (2016) and Zhang et al. (2015) have successfully applied convolutional neural networks to model implicit relations within the PDTB-framework. Along these lines and inspired by the work in Weiss (2015), we also see great potential in the use of neural network-based techniques to SDP. Similarly, our approach trains a modular component for shallow discourse parsing which incorporates distributed word representations for argument spans by abstraction from surface-level (token) information. Crucially, our

approach substitutes the traditional sparse and hand-crafted features from the literature to account for a minimalist, but at the same time, general (latent) representation of the discourse units. In the next sections, we elaborate on our novel neural network-based approach for implicit sense labeling and how it is fit into the overall system architecture of the parser.

3 A Neural Sense Labeler for Implicit and Entity Relations

We construct a neural network-based module for the classification of senses for both implicit and entity (*EntRel*) relations.[3] As a very general and highly data-driven approach to modeling discourse relations, our classifier incorporates *only* word embeddings and basic syntactic dependency information. Also, in order to keep the setup easily adaptable to new data and other languages, we avoid the use of very specific and costly hand-crafted features (such as sentiment polarities, word-pair features, cue phrases, modality, production rules, highly specific semantic information from external ontologies such as VerbNet, etc.), which has been the main focus in traditional approaches to SDP (Huang and Chen, 2011; Park and Cardie, 2012; Feng and Hirst, 2012). Instead, we substitute (sparse) tokens in the argument spans, with dense, distributed representations, i.e. word embeddings, as the main source of information for the sense classification component. Closely related, Zhang et al. (2015) have explored a similar approach of constructing argument vectors by applying a set of aggregation functions on their token vectors, however, without the use of additional (syntactic) information, while embedding their vectors into a single-layer neural network only.

In our experiments, we used the pre-trained *GoogleNews* vectors (for English) and the *Gigaword*-induced vectors (for Chinese) provided by the shared task as a starting point.[4] We further trained the word vectors on the raw Wall Street Journal texts, thus tuning the embeddings toward the data at hand, with the goal of considerably improving their predictive power in the sense classification task. Specifically, the pre-trained vectors of size 300 were updated by the skip-gram method (Mikolov et al., 2013)[5] in multiple passes over the Newswire texts with decreasing learning rate. This procedure is supposed to improve the quality of the embeddings and also their coverage.

Our new word vector model provides general vector representations for each token in the two argument spans[6], which forms the basis for producing compositional vectors to represent the two spans. Compositional vectors that introduce a fixed-length representation of a variable-length span of tokens are practical features for feedforward neural networks. Thus, we may combine the token vectors of each span by simply averaging vectors, or – following Mitchell and Lapata (2008) – by calculating an aggregated argument vector $\vec{v'}$:

$$\vec{v'}(j) = \frac{1}{k(j)} \sum_{i=1}^{k(j)} V(j)_i + \prod_{i=1}^{k(j)} V(j)_i \quad (1)$$

for arguments $j \in \{1, 2\}$, where $k(j) = |t(j)|$ defines their lengths in the number of tokens and $\prod$ applies the pointwise product $\odot$ over the token vectors in $V(j)$.

Both procedures produce rather simple argument representations that do not account for word order variation or any other sentence structure information, yet they serve as decent features for discourse parsing and other related tasks. By introducing pointwise multiplication of the token vectors, the elements that represent assumed independent, latent semantic dimensions are not merely lumped together across vectors, but are allowed to scale according to their mutual relevance.[7]

Improving upon the compositional representation produced by Equation 1, we incorporate additional syntactic dependency information: for each token in an argument span, we calculate the depth d from the corresponding sentence's root node and weight the token vector by $\frac{1}{2^d}$ before applying the

[3] The reason to combine both relation types has been a design decision as *EntRels* are very similar to implicit relations and are also missing a connective. *AltLex* relations seemed too few to have any statistical impact on the performance of our experiments and have been ignored altogether.

[4] `http://www.cs.brandeis.edu/~clp/conll16st/dataset.html`

[5] We found window size of 8 and min term count = 3 to be optimal. Neural networks were trained using the *gensim* package: `http://radimrehurek.com/gensim/`.

[6] We ignore unknown tokens for which no vectors exist.

[7] In our experiments, Equation 1 outperformed simpler strategies of either average or multiplication alone. This also indicates that it is beneficial to not completely suppress dimensions with near-zero values for single tokens.

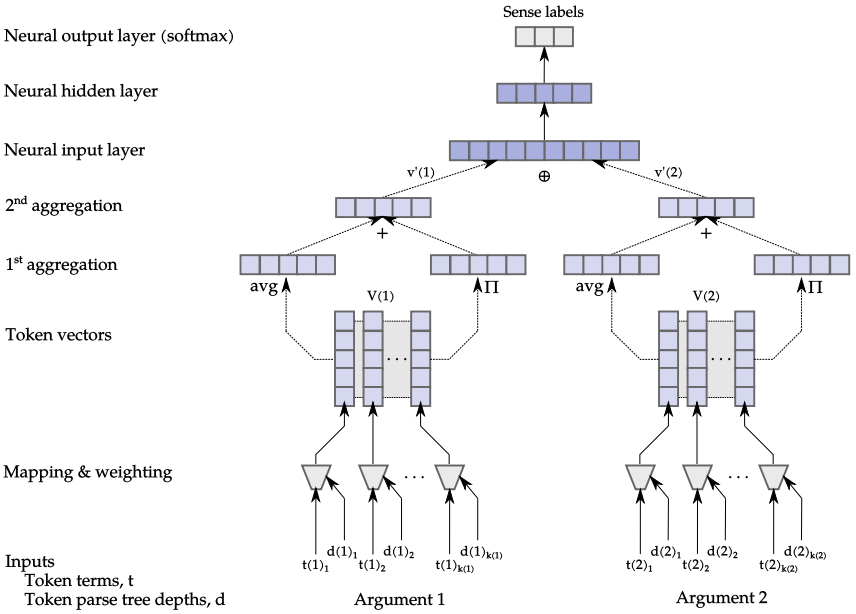


Figure 1: The feature construction process from argument spans (light blue) and neural architecture (dark blue) for implicit sense classification (incl. *EntRel*) . Dotted lines represent pointwise vector operations.

aggregating operators.[8]

The bottom of Figure 1 illustrates the first step of the process, i.e. mapping tokens to their corresponding vectors based on the updated word vector model, as well as the token depth weighting. Secondly, the aggregation operators are applied, i.e., the sum ($+$) of the pointwise product ($\prod/\odot$) and average (*avg*) of the vectors. Finally, the compositional vectors for each of the arguments are concatenated ($\oplus$) and serve as input to a feedforward neural network.

Given the composed argument vectors, we set up a network with one hidden layer and a softmax output layer to classify among 20 implicit senses for English and 9 for Chinese, plus an additional *EntRel* label. Other relations, such as *AltLex*, are not modeled. We train the network using Nesterov's Accelerated Gradient (Nesterov, 1983) and optimized all hyper-parameters on the development set. Best results were achieved with *rectified linear activation with learnable leak rate and gain (lgrelu)*, 40-60 hidden nodes and weight decay and hidden node regularization of 0.0001.[9]

4 The Competition Tasks & Pipelines

We participate in the *closed track* of the shared task, specifically in both *full* and *supplementary tasks* (*sense-only*) on English and Chinese texts. Full tasks require a participant's system to identify argument pairs and to label the sense relation that holds between them. In each supplementary task, gold arguments are provided so that the performance of sense labeling does not suffer from error propagation due to incorrectly detected argument spans.

We combine different *existent* modules to address the specific settings and classification needs of both full and supplementary tasks for both lan-

[8] Tokens that are missing in the parse tree, such as punctuation symbols, are weighted by 0.25, in our optimal setting.

[9] The learning rate was set to 0.0001. Momentum of 0.35-0.6 and 60 hidden nodes performed well for the English tasks, and momentum of 0.85 and 40 hidden nodes for Chinese (with fewer output nodes). Good results were also obtained by *Parametric Rectified Linear Unit (prelu)* activation, as well as the combination of larger hidden layer and stronger regularization (e.g., L1 regularization of 0.1 on 100 nodes).

guages. The modules and their combination with our implicit neural sense classifier will be outlined in the following sections.

4.1 English Full Task Pipeline (EFTP)

For the full task, we exploit the high-quality argument extraction modules of the two best-performing systems by Wang and Lan (2015, W&L) and Stepanov et al. (2015) from last year's competition (re-using their original implementations): Specifically, we initially run both systems for all *explicit* relations only, and keep those predicted arguments and sense labels – from either of the two systems – which maximize F_1-score on the development set. With this simple heuristic, we hope to improve upon the best results from W&L, as, for instance, Stepanov et al. (2015) perform particularly well on all temporal relations, while W&L's tool handles the majority of other senses well.

For all implicit and *EntRel* relations, we keep the exact argument spans obtained from the W&L system and reject all sense labels. In a second step, we *re-classify* all these implicit relations by our neural net-based architecture described in Section 3 given only the tokens and their dependencies in both argument spans. Finally, we merge all combined explicit and re-classified implicit relations into the final set for evaluation.

4.2 English Supplementary Task Pipeline (ESTP)

We make use of the system by Stepanov et al. (2015) to label all *explicit* relation senses, and classify all other relations with an empty token list for connectors (i.e., implicit and *EntRels*) by our neural network architecture from Section 3.

4.3 Chinese Full Task Pipeline (CFTP)

Since for the Chinese full task no reusable argument extraction tools were available, we have set up a minimalist (baseline) implementation whose individual steps we sketch briefly:

1. **Connective detection** is realized by means of a sequence labeling/CRF model.[10] Features are unigram and bigram information from the tokens, their parts-of-speech, dependency head, dependency chain, whether the token is found as a connector in the training set, and its relative position within the sentence.
2. **Argument extraction** is based on the output of predicted connectives for both inter- and intra-sentence relations. As an additional feature, we found the IOB chain for the syntactic path of a token to be useful.[11]
3. We heuristically **post-process** the CRF-labeled argument tokens in order to assign connectors to same-sentence or separate-sentence *Arg1* and *Arg2* spans.
4. The so-obtained **explicit argument pairs** are sense labeled by a (linear-kernel) SVM classifier[12] with the connector word as the only feature, following the minimalist setting in Chiarcos and Schenk (2015).
5. As **implicit relations** we consider *all inter-sentential relations* which are not already part of an explicit relation. Same-sentence relations are ignored altogether.

4.4 Chinese Supplementary Task Pipeline (CSTP)

For the provided argument pairs, we label *explicit* relations (i.e. those containing a non-empty connector) by the SVM classifier which has been trained using only a single feature – the connector token. For all other relations, we again employ our neural network-based strategy described in Section 3. The overall architecture is exactly the same as for the English subtask; only the (hyper)parameters have been updated in accordance with the Chinese training data.

5 Evaluation

5.1 English Full Task

Table 1 shows the performance of our full-task pipeline (EFTP) which integrates our novel feed-forward neural network architecture for implicit sense labeling. The figures suggest that our minimalist approach is highly competitive and can even outperform the best results from last year's competition in terms of F_1-scores on two out of three evaluation sets (cf. last *implicit* column).

Overall, with the integration of the combined systems by W&L and Stepanov et al. (2015), we can improve upon the state-of-the-art by an absolute increase in F_1-score of 0.5% on the blind test

[10] https://taku910.github.io/crfpp/

[11] This information was generated using the script from http://ilk.uvt.nl/team/sabine/chunklink/chunklink_2-2-2000_for_conll.pl

[12] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

set– which is marginal but only due to the fruitful re-classification of the already-provided (and therefore fixed) argument spans.

Measured on the development set, we found that the *dependency depth weighting* contributes to an absolute improvement in accuracy of 1.5% for non-explicit relations.

set	system	overall	explicit	*implicit*
dev	W&L	37.84	48.16	28.70
	EFTP	**40.21**	**50.87**	**30.99**
test	W&L	29.69	39.96	**20.74**
	EFTP	**29.78**	**40.44**	20.60
blind	W&L	24.00	30.38	18.78
	EFTP	**24.47**	**30.74**	**19.63**

Table 1: English full task F_1-scores.

5.2 English Supplementary Task

Without error propagation from argument identification, and with the gold arguments provided in the evaluation sets, the performance of our implicit sense labeling component is even better; cf. Table 2: on both PDTB evaluation sets F_1-scores increase by 2.7% and 3.16% (absolute) and by 6.32% and up to **9.17%** (relative) on the development and test section, respectively.

Strikingly, however, the prediction quality on the blind test set is worse than expected. We assume that this is partly due to the (slightly) heterogeneous content of the annotated *Wikinews*, as opposed to the original Penn Discourse Treebank data on which our system performs extraordinarily well.

set	system	overall	explicit	*implicit*
dev	W&L	65.11	90.00	42.72
	ESTP	**66.90**	**91.35**	**45.42**
test	W&L	61.27	**90.79**	34.45
	ESTP	**62.64**	90.13	**37.61**
blind	W&L	**54.76**	**76.44**	**36.29**
	ESTP	52.32	76.40	31.85

Table 2: English sense-only task F_1-scores.

5.3 Chinese Full Task

This year's edition of the shared task has been the first to address shallow discourse parsing for Chinese Newswire texts. Given no prior (directly comparable) results on Chinese SDP so far, we simply report the performance of our system on all evaluation sets in Table 3.

set	system	overall	explicit	*implicit*
dev	**CFTP**	22.16	17.45	**22.67**
test	**CFTP**	**24.21**	**28.73**	22.26
blind	**CFTP**	12.90	18.56	10.80

Table 3: Chinese full task F_1-scores.

5.4 Chinese Supplementary Task

A final evaluation has been concerned with the sense-only labeling of gold-provided arguments for Chinese. We want to point out that the neural network architecture for implicit relations (with 70.59% F_1-score on the dev set, cf. Table 4) has beaten all our other experiments: In particular, we have conducted an SVM setup in which we employed the traditional word-pair features substituted by Brown clusters 3200 (65.12%), and special additive Arg1/Arg2 combinations of word embeddings – yielding only 62.8% which equals the majority class baseline indicating no predictive power for any given kernel type.

set	system	overall	explicit	*implicit*
dev	**CSTP**	75.72	96.10	70.59
test	**CSTP**	**77.01**	**96.34**	**71.87**
blind	**CSTP**	63.73	80.39	57.59

Table 4: Chinese sense-only task F_1-scores.

6 Conclusion

In the context of the CoNLL 2016 Shared Task on shallow discourse parsing, we have described our participating system and its architecture. Specifically, we introduced a novel feedforward neural network-based component for implicit sense labeling whose only source of information are pretrained word embeddings and syntactic dependencies. Its highly generic and extremely simple design is the main advantage of this module. It has proven to be language-independent, easy to tune and optimize and does not require the use of handcrafted – rich – linguistic features.

Still its performance is highly competitive with the state-of-the-art on implicit sense labeling and builds a solid groundwork for future extensions.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August.

Or Biran and Kathleen McKeown. 2013. Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 69–73.

Or Biran and Kathleen McKeown. 2015. PDTB Discourse Parsing as a Tagging Task: The Two Taggers Approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104, Prague, Czech Republic, September. Association for Computational Linguistics.

Christian Chiarcos and Niko Schenk. 2015. A Minimalist Approach to Shallow Discourse Parsing and Implicit Relation Recognition. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 42–49.

David A. duVerle and Helmut Prendinger. 2009. A Novel Discourse Parser Based on Support Vector Machine Classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 665–673, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sucheta Ghosh, Giuseppe Riccardi, and Richard Johansson. 2012. Global Features for Shallow Discourse Parsing. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 150–159.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation Learning for Text-level Discourse Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland, June. Association for Computational Linguistics.

Fang Kong, Tou Hwee Ng, and Guodong Zhou. 2014. A Constituent-Based Approach to Argument Labeling with Joint Inference in Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 68–77. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive Deep Models for Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar, October. Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 343–351, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184, 4.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit Discourse Relation Classification via Multi-Task Neural Networks. *CoRR*, abs/1603.02776.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *Proceedings of Association for Computational Linguistics*, pages 236–244.

Yurii Nesterov. 1983. A method of solving a convex programming problem with convergence rate O (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376.

Tsuyoshi Okita, Longyue Wang, and Qun Liu. 2015. The DCU Discourse Parser: A Sense Classification Task. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 71–77, Beijing, China, July. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2012. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 108–112, Seoul, South Korea, July. Association for Computational Linguistics, Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 13–16.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 683–691, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September. Springer.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *In Proceedings of LREC*.

Attapol Rutherford and Nianwen Xue. 2014. Discovering Implicit Discourse Relations Through Brown Cluster Pair Representation and Coreference Patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

Evgeny Stepanov, Giuseppe Riccardi, and Orkan Ali Bayer. 2015. The UniTN Discourse Parser in CoNLL 2015 Shared Task: Token-level Sequence Labeling with Argument-specific Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 25–31. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2015. A Refined End-to-End Discourse Parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24. Association for Computational Linguistics.

Longyue Wang, Chris Hokamp, Tsuyoshi Okita, Xiaojun Zhang, and Qun Liu. 2015. The DCU Discourse Parser for Connective, Argument Identification and Explicit Sense Classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 89–94. Association for Computational Linguistics.

Bonnie L. Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779.

Gregor Weiss. 2015. Learning Representations for Text-level Discourse Parsing. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 16–21, Beijing, China, July. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning: Shared Task*, Beijing, China.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2230–2235.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 69–77, Jeju Island, Korea, July. Association for Computational Linguistics.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting Discourse Connectives for Implicit Discourse Relation Recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1507–1514, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paper VI

VI

A recurrent neural model with attention for the recognition of Chinese implicit discourse relations

S. Rönnqvist and N. Schenk and C. Chiarcos (2017). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 256–262. Association for Computational Linguistics

A Recurrent Neural Model with Attention for the Recognition of Chinese Implicit Discourse Relations

Samuel Rönnqvist[1,2,*], **Niko Schenk**[2,*] and **Christian Chiarcos**[2]
[1]Turku Centre for Computer Science – TUCS, Åbo Akademi University, Turku, Finland
[2]Applied Computational Linguistics Lab, Goethe University, Frankfurt am Main, Germany
sronnqvi@abo.fi
{schenk,chiarcos}@informatik.uni-frankfurt.de

Abstract

We introduce an attention-based Bi-LSTM for Chinese implicit discourse relations and demonstrate that modeling argument pairs as a joint sequence can outperform word order-agnostic approaches. Our model benefits from a partial sampling scheme and is conceptually simple, yet achieves state-of-the-art performance on the Chinese Discourse Treebank. We also visualize its attention activity to illustrate the model's ability to selectively focus on the relevant parts of an input sequence.

1 Introduction

True text understanding is one of the key goals in Natural Language Processing and requires capabilities beyond the lexical semantics of individual words or phrases. Natural language descriptions are typically driven by an inter-sentential coherent structure, exhibiting specific *discourse* properties, which in turn contribute significantly to the global meaning of a text. Automatically detecting how meaning units are organized benefits practical downstream applications, such as question answering (Sun and Chai, 2007), recognizing textual entailment (Hickl, 2008), sentiment analysis (Trivedi and Eisenstein, 2013), or text summarization (Hirao et al., 2013).

Various formalisms in terms of semantic coherence frameworks have been proposed to account for these contextual assumptions (Mann and Thompson, 1988; Lascarides and Asher, 1993; Webber, 2004). The annotation schemata of the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB), for instance, define discourse units as syntactically motivated character spans in the text, augmented with relations pointing from the second argument (*Arg2*, prototypically, a discourse unit associated with an explicit discourse marker) to its antecedent, i.e., the discourse unit *Arg1*. Relations are labeled with a relation type (its sense) and the associated discourse marker. Both, PDTB and CDTB, distinguish *explicit* from *implicit* relations depending on the presence of such a marker (e.g., *because*/ 因).[1] Sense classification for implicit relations is by far more challenging because the argument pairs lack the marker as an important feature. Consider, for instance, the following example from the CDTB as implicit CONJUNCTION:

Arg1: 会谈就一些原则和具体问题进行了深入讨论，达成了一些谅解 *In the talks, they discussed some principles and specific questions in depth, and reached some understandings*

Arg2: 双方一致认为会谈具有积极成果 *Both sides agree that the talks have positive results*

Motivation: Previous work on implicit sense labeling is heavily feature-rich and requires domain-specific, semantic lexicons (Pitler et al., 2009; Feng and Hirst, 2012; Huang and Chen, 2011). Only recently, resource-lean architectures have been proposed. These promising neural methods attempt to infer latent representations appropriate for implicit relation classification (Zhang et al., 2015; Ji et al., 2016; Chen et al., 2016). So far, unfortunately, these models have been evaluated *only* on four top-level senses—sometimes even with inconsistent evaluation setups.[2] Furthermore, most systems have initially been designed for the English PDTB and involve complex, task-

[*]Both first authors contributed equally to this work.

[1]The set of relation types and senses is completed by alternative lexicalizations (ALTLEX/discourse marker rephrased), and entity relations (ENTREL/anaphoric coherence).

[2]E.g., four binary classifiers vs. four-way classification.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 256–262

https://doi.org/10.18653/v1/P17-2040

specific architectures (Liu and Li, 2016), while discourse modeling techniques for Chinese have received very little attention in the literature and are still seriously underrepresented in terms of publicly available systems. What is more, over 80% of all words in Chinese discourse relations are implicit—compared to only 52% in English (Zhou and Xue, 2012).

Recently, in the context of the CoNLL 2016 shared task (Xue et al., 2016), a first independent evaluation platform beyond class level has been established. Surprisingly, the best performing neural architectures to date are standard *feedforward* networks, cf. Wang and Lan (2016); Schenk et al. (2016); Qin et al. (2016). Even though these specific models completely ignore word order within arguments, such feedforward architectures have been claimed by Rutherford et al. (2016) to generally outperform any thoroughly-tuned recurrent architecture.

Our Contribution: In this work, we release the first attention-based *recurrent* neural sense classifier, specifically developed for Chinese implicit discourse relations. Inspired by Zhou et al. (2016), our system is a practical adaptation of the recent advances in relation modeling extended by a novel sampling scheme.

Contrary to previous assertions by Rutherford et al. (2016), our model demonstrates superior performance over traditional bag-of-words approaches with feedfoward networks by treating discourse arguments as a joint sequence. We evaluate our method within an independent framework and show that it performs very well beyond standard class-level predictions, achieving state-of-the-art accuracy on the CDTB test set.

We illustrate how our model's attention mechanism provides means to highlight those parts of an input sequence that are relevant for the classification decision, and thus, it may enable a better understanding of the implicit discourse parsing problem. Our proposed network architecture is flexible and largely language-independent as it operates only on word embeddings. It stands out due to its structural simplicity and builds a solid ground for further development towards other textual domains.

2 Approach

We propose the use of an attention-based bidirectional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997, LSTM) network to predict senses of discourse relations. The model draws upon previous work on LSTM, in particular its bidirectional mode of operation (Graves and Schmidhuber, 2005), attention mechanisms for recurrent models (Bahdanau et al., 2014; Hermann et al., 2015), and the combined use of these techniques for entity relation recognition in annotated sequences (Zhou et al., 2016). More specifically, our model is a flexible recurrent neural network with capabilities to *sequentially* inspect tokens and to highlight which parts of the input sequence are most informative for the discourse relation recognition task, using the weighting provided by the attention mechanism. Furthermore, the model benefits from a novel sampling scheme for arguments, as elaborated below. The system is learned in an end-to-end manner and consists of multiple layers, which are illustrated in Figure 1.

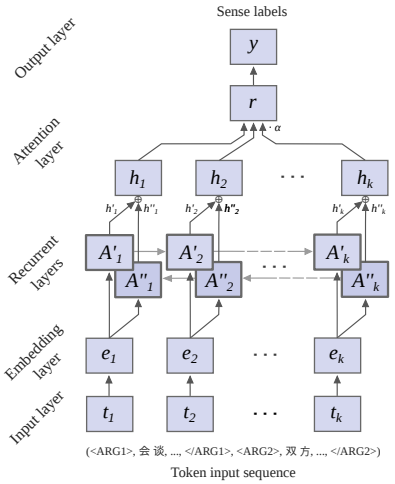


Figure 1: The attention-based bidirectional LSTM network for the task of modeling argument pairs for Chinese implicit discourse relations.

First, token sequences are taken as input and special markers (<ARG1>, </ARG1>, etc.) are inserted into the corresponding positions to inform the model on the start and end points of argument spans. This way, we can ensure a general flexibility in modeling discourse units and could easily extend them with additional context, for instance. In our experiments on implicit arguments,

only the tokens in the respective spans are considered. Note that, unlike previous works, our approach models *Arg1-Arg2* pairs as a *joint* sequence and does not first compute intermediate representations of arguments separately.

Second, an input layer encodes tokens using one-hot vector representations (t_i for tokens at positions $i \in [1, k]$), and a subsequent embedding layer provides a dense representation (e_i) to serve as input for the recurrent layers. The embedding layer is initialized using pre-trained word vectors, in our case 300-dimensional Chinese Gigaword vectors (Graff and Chen, 2005).[3] These embeddings are further tuned as the network is trained towards the prediction task. Embeddings for unknown tokens, e.g., markers, are trained by back-propagation only. Note that, tokens, markers and the pre-trained vectors represent the only source of information for the prediction task.

For the recurrent setup, we use a layer of LSTM networks in a bidirectional manner, in order to better capture dependencies between parts of the input sequence by inspection of both left and right-hand-side contexts at each time step. The LSTM holds a state representation as a continuous vector passed to the subsequent time step, and it is capable of modeling long-range dependencies due to its gated memory. The forward (A') and backward (A'') LSTMs traverse the sequence e_i, producing sequences of vectors h'_i and h''_i respectively, which are then summed together (indicated by $\oplus$ in Figure 1).

The resulting sequence of vectors h_i is reduced into a single vector and fed to the final softmax output layer in order to classify the sense label y of the discourse relation. This vector may be obtained either as the final vector h produced by an LSTM, or through pooling of all h_i, or by using attention, i.e., as a weighted sum over h_i. While the model may be somewhat more difficult to optimize using attention, it provides the added benefit of interpretability, as the weights highlight to what extent the classifier considers the LSTM state vectors at each token during modeling. This is particularly interesting for discourse parsing, as most previous approaches have provided little support for pinpointing the driving features in each argument span.

Finally, the attention layer contains the trainable vector w (of the same dimensionality as vectors h_i) which is used to dynamically produce a weight vector α over time steps i by:

$$\alpha = softmax(w^T tanh(H))$$

where H is a matrix consisting of vectors h_i. The output layer r is the weighted sum of vectors in H:

$$r = H\alpha^T$$

Partial Argument Sampling: For the purpose of enlarging the instance space of training items in the CDTB, and thus, in order to improve the predictive performance of the model, we propose a novel *partial sampling* scheme of arguments, whereby the model is trained and validated on sequences containing both arguments, as well as *single* arguments. A data point (a_1, a_2, y), with a_i being the token sequence of argument i, is expanded into $\{(a_1, a_2, y), (a_1, a_2, y), (a_1, y), (a_2, y)\}$. We duplicate bi-argument samples (a_1, a_2, y) (in training and development data only) to balance their frequencies against single-argument samples.

Two lines of motivation support the inclusion of single argument training examples, grounded in linguistics and machine learning, respectively. First, it has been shown that single arguments in isolation can evoke a strong expectation towards a certain implicit discourse relation, cf. Asr and Demberg (2015) and, in particular, Rohde and Horton (2010) in their psycholinguistic study on *implicit causality verbs*. Second, the procedure may encourage the model to learn better representations of individual argument spans in support of modeling of arguments in composition, cf. LeCun et al. (2015). Due to these aspects, we believe this data augmentation technique to be effective in reinforcing the overall robustness of our model.

Implementational Details: We train the model using fixed-length sequences of 256 tokens with zero padding at the beginning of shorter sequences and truncate longer ones. Each LSTM has a vector dimensionality of 300, matching the embedding size. The model is regularized by 0.5 dropout rate between the layers and weight decay ($2.5e^{-6}$) on the LSTM inputs. We employ Adam optimization (Kingma and Ba, 2014) using the cross-entropy loss function with mini batch size of 80.[4]

[3] http://www.cs.brandeis.edu/~clp/conll16st/dataset.html

[4] The model is implemented in *Keras* https://keras.io/.

CDTB Development Set			CDTB Test Set		
Rank	System	% accuracy	Rank	System	% accuracy
1	Wang and Lan (2016)	73.53	1	Wang and Lan (2016)	72.42
2	Qin et al. (2016)	71.57	2	Schenk et al. (2016)	71.87
3	Schenk et al. (2016)	70.59	3	Rutherford and Xue (2016)	70.47
4	Rutherford and Xue (2016)	68.30	4	Qin et al. (2016)	67.41
5	Weiss and Bajec (2016)	66.67	5	Weiss and Bajec (2016)	64.07
6	Weiss and Bajec (2016)	61.44	6	Weiss and Bajec (2016)	63.51
7	Jian et al. (2016)	21.90	7	Jian et al. (2016)	21.73
	This Paper:	**93.52***		**This Paper:**	**73.01**

Table 1: Non-explicit parser scores on the official CoNLL 2016 CDTB development and test sets. (*Scores on development set are obtained through partial sampling and are not directly comparable.)

Sense Label	Training	Dev't	Test
CONJUNCTION	5,174	189	228
majority class	(66.3%)	(62.8%)	(64.8%)
EXPANSION	1,188	48	40
ENTREL	1,099	50	71
CAUSATION	187	10	8
CONTRAST	66	3	1
PURPOSE	56	1	3
CONDITIONAL	26	0	1
TEMPORAL	26	0	0
PROGRESSION	7	0	0
# impl. rels	7,804	301	352

Table 2: Implicit sense labels in the CDTB.

3 Evaluation

We evaluate our recurrent model on the CoNLL 2016 shared task data[5] which include the official training, development and test sets of the CDTB; cf. Table 2 for an overview of the implicit sense distribution.[6]

In accordance with previous setups (Rutherford et al., 2016), we treat entity relations (ENTREL) as implicit and exclude ALTLEX relations. In the evaluation, we focus on the *sense-only* track, the subtask for which gold arguments are provided and a system is supposed to label a given argument pair with the correct sense. The results are shown in Table 1.

With our proposed architecture it is possible to correctly label 257/352 (73.01%) of implicit relations on the test set, outperforming the best feed-forward system of Wang and Lan (2016) and all other word order-agnostic approaches. Development and test set performances suggest the robustness of our approach and its ability to generalize to unseen data.

Ablation Study: We perform an ablation study to quantitatively assess the contribution of two of the characteristic aspects of our model. First, we compare the use of the attention mechanism against the simpler alternative of feeding the final LSTM hidden vectors (h'_k and h''_1) directly to the output layer. When attention is turned off, this yields an absolute decrease in performance of 2.70% on the test set, which is substantial and significant according to a Welch two-sample t-test ($p < .001$). Second, we independently compare the use of the partial sampling scheme against training on the standard argument pairs in the CDTB. Here, the absence of the partial sampling scheme yields an absolute decrease in accuracy of 5.74% ($p < .001$), which demonstrates its importance for achieving competitive performance on the task.

Performance on the PDTB: As a side experiment, we investigate the model's language independence by applying it to the implicit argument pairs of the English PDTB. Due to computational time constraints we do not optimize hyperparameters, but instead train the model using identical settings as for Chinese, which is expected to lead to suboptimal performance on the evaluation data. Nevertheless, we measure 27.09% accuracy on the PDTB test set (surpassing the majority class baseline of 22.01%), which shows that the model has potential to generalize across implicit discourse relations in a different language.

[5] http://www.cs.brandeis.edu/~clp/conll16st/

[6] Note that, in the CDTB, implicit relations appear almost *three times more often* than explicit relations. Out of these, 65% appear within the same sentence. Finally, 25 relations in the training set have two labels.

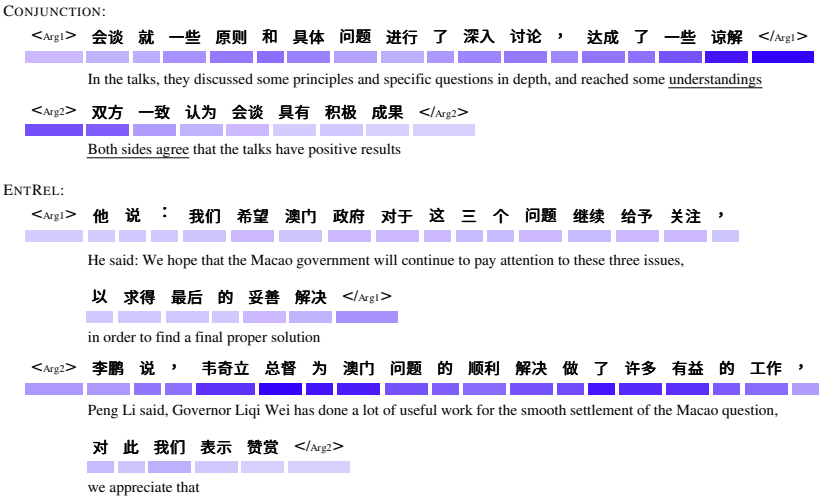


Figure 2: Visualization of attention weights for Chinese characters with high (dark blue) and low (light blue) intensities. The underlined English phrases are semantically structure-shared by the two arguments.

Visualizing Attention Weights: Finally, in Figure 2, we illustrate the learned attention weights which pinpoint important subcomponents within a given implicit discourse relation. For the implicit CONJUNCTION relation the weights indicate a peak on the transition between the argument boundary, establishing a connection between the semantically related terms *understandings–agree*. Most ENTRELs show an opposite trend: here second arguments exhibit larger intensities than *Arg1*, as most entity relations follow the characteristic writing style of newspapers by adding additional information by reference to the same entity.

4 Summary & Outlook

In this work, we have presented the first attention-based recurrent neural sense labeler specifically developed for Chinese implicit discourse relations. Its ability to model discourse units sequentially and jointly has been shown to be highly beneficial, both in terms of state-of-the-art performance on the CDTB (outperforming word order-agnostic feedforward approaches), and also in terms of insightful observations into the inner workings of the model through its attention mechanism. The architecture is structurally simple, benefits from partial argument sampling, and can be easily adapted to similar relation recognition tasks. In future work, we intend to extend our approach to different languages and domains, e.g., to the recent data sets on narrative story understanding or question answering (Mostafazadeh et al., 2016; Feng et al., 2015). We believe that recurrent modeling of implicit discourse information can be a driving force in successfully handling such complex semantic processing tasks.[7]

Acknowledgments

The authors would like to thank Ayah Zirikly, Philip Schulz and Wei Ding for their very helpful suggestions on an early draft version of the paper, and also thank the anonymous reviewers for their valuable feedback and insightful comments. We are grateful to Farrokh Mehryary for technical support with the attention layer implementation. Computational resources were provided by CSC – IT Centre for Science, Finland, and Arcada University of Applied Sciences, Helsinki, Finland. Our research at Goethe University Frankfurt was supported by the project 'Linked Open Dictionaries (LiODi, 2015-2020)', funded by the German Ministry for Education and Research (BMBF).

[7]The code involved in this study is publicly available at http://www.acoli.informatik.uni-frankfurt.de/resources/.

References

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform Information Density at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission. In *11th International Conference on Computational Semantics (IWCS)*. page 118. http://www.coli.uni-saarland.de/ fatemeh/iwcs2015.pdf.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. http://aclweb.org/anthology/P/P16/P16-1163.pdf.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*. pages 813–820. https://doi.org/10.1109/ASRU.2015.7404872.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 60–68. http://www.aclweb.org/anthology/P12-1007.

David Graff and Ke Chen. 2005. Chinese Gigaword. LDC Catalog No.: LDC2003T09, ISBN, 1:58563–58230.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5-6):602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. pages 1693–1701.

Andrew Hickl. 2008. Using Discourse Commitments to Recognize Textual Entailment. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '08, pages 337–344. http://dl.acm.org/citation.cfm?id=1599081.1599124.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-Document Summarization as a Tree Knapsack Problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1515–1520. http://aclweb.org/anthology/D/D13/D13-1158.pdf.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, pages 1442–1446. http://www.aclweb.org/anthology/I11-1170.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A Latent Variable Recurrent Neural Network for Discourse-Driven Language Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 332–342. http://www.aclweb.org/anthology/N16-1037.

Ping Jian, Xiaohan She, Chenwei Zhang, Pengcheng Zhang, and Jian Feng. 2016. Discourse Relation Sense Classification Systems for CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 158–163. https://doi.org/10.18653/v1/K16-2022.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Alex Lascarides and Nicholas Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense entailment. *Linguistics and Philosophy* 16(5):437–493.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553):436–444.

Yang Liu and Sujian Li. 2016. Recognizing Implicit Discourse Relations via Repeated Reading: Neural Networks with Multi-Level Attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 1224–1233. http://aclweb.org/anthology/D/D16/D16-1130.pdf.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3):243–281.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 839–849. http://www.aclweb.org/anthology/N16-1098.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL 2009, pages 683–691. http://www.aclweb.org/anthology/P/P09/P09-1077.pdf.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, pages 2961–2968.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow Discourse Parsing Using Convolutional Neural Network. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 70–77. https://doi.org/10.18653/v1/K16-2010.

Hannah Rohde and William Horton. 2010. Why or what next? Eye movements reveal expectations about discourse direction. Talk at the 23rd Annual CUNY Conference on Human Sentence Processing. New York, NY.

Attapol Rutherford and Nianwen Xue. 2016. Robust Non-Explicit Neural Discourse Parser in English and Chinese. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 55–59. https://doi.org/10.18653/v1/K16-2007.

Attapol T. Rutherford, Vera Demberg, and Nianwen Xue. 2016. Neural Network Models for Implicit Discourse Relation Classification in English and Chinese without Surface Features. *CoRR* abs/1606.01990. http://arxiv.org/abs/1606.01990.

Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Rönnqvist, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Do We Really Need All Those Rich Linguistic Features? A Neural Network-Based Approach to Implicit Sense Labeling. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 41–49. https://doi.org/10.18653/v1/K16-2005.

Mingyu Sun and Joyce Y Chai. 2007. Discourse processing for context question answering based on linguistic knowledge. *Knowledge-Based Systems* 20(6):511–526.

Rakshit S. Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. pages 808–813. http://aclweb.org/anthology/N/N13/N13-1100.pdf.

Jianxiang Wang and Man Lan. 2016. Two End-to-end Shallow Discourse Parsers for English and Chinese in CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 33–40. https://doi.org/10.18653/v1/K16-2004.

Bonnie L. Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28(5):751–779. http://dblp.uni-trier.de/db/journals/cogsci/cogsci28.html.

Gregor Weiss and Marko Bajec. 2016. Discourse Sense Classification from Scratch using Focused RNNs. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 50–54. https://doi.org/10.18653/v1/K16-2006.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*. Association for Computational Linguistics, Berlin, Germany.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 2230–2235. http://aclweb.org/anthology/D/D15/D15-1266.pdf.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. http://aclweb.org/anthology/P/P16/P16-2034.pdf.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, pages 69–77. http://www.aclweb.org/anthology-new/P/P12/P12-1008.bib.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases
202. **Fredrik Abbors**, Model-Based Testing of Software Systems: Functionality and Performance
203. **Inna Pereverzeva**, Formal Development of Resilient Distributed Systems
204. **Mikhail Barash**, Defining Contexts in Context-Free Grammars
205. **Sepinoud Azimi**, Computational Models for and from Biology: Simple Gene Assembly and Reaction Systems
206. **Petter Sandvik**, Formal Modelling for Digital Media Distribution

207. **Jongyun Moon**, Hydrogen Sensor Application of Anodic Titanium Oxide Nanostructures
208. **Simon Holmbacka**, Energy Aware Software for Many-Core Systems
209. **Charalampos Zinoviadis**, Hierarchy and Expansiveness in Two-Dimensional Subshifts of Finite Type
210. **Mika Murtojärvi**, Efficient Algorithms for Coastal Geographic Problems
211. **Sami Mäkelä**, Cohesion Metrics for Improving Software Quality
212. **Eyal Eshet**, Examining Human-Centered Design Practice in the Mobile Apps Era
213. **Jetro Vesti**, Rich Words and Balanced Words
214. **Jarkko Peltomäki**, Privileged Words and Sturmian Words
215. **Fahimeh Farahnakian**, Energy and Performance Management of Virtual Machines: Provisioning, Placement and Consolidation
216. **Diana-Elena Gratie**, Refinement of Biomodels Using Petri Nets
217. **Harri Merisaari**, Algorithmic Analysis Techniques for Molecular Imaging
218. **Stefan Grönroos**, Efficient and Low-Cost Software Defined Radio on Commodity Hardware
219. **Noora Nieminen**, Garbling Schemes and Applications
220. **Ville Taajamaa**, O-CDIO: Engineering Education Framework with Embedded Design Thinking Methods
221. **Johannes Holvitie**, Technical Debt in Software Development – Examining Premises and Overcoming Implementation for Efficient Management
222. **Tewodros Deneke**, Proactive Management of Video Transcoding Services
223. **Kashif Javed**, Model-Driven Development and Verification of Fault Tolerant Systems
224. **Pekka Naula**, Sparse Predictive Modeling – A Cost-Effective Perspective
225. **Antti Hakkala**, On Security and Privacy for Networked Information Society – Observations and Solutions for Security Engineering and Trust Building in Advanced Societal Processes
226. **Anne-Maarit Majanoja**, Selective Outsourcing in Global IT Services – Operational Level Challenges and Opportunities
227. **Samuel Rönnqvist**, Knowledge-Lean Text Mining

Turku
Centre *for*
Computer
Science

http://www. tucs.fi

tucs@abo.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3622-8
ISSN 1239-1883

Samuel Rönnqvist

Knowledge-Lean Text Mining

Samuel Rönnqvist

Knowledge-Lean Text Mining

Samuel Rönnqvist

Knowledge-Lean Text Mining